ED 370 436 FL 022 217

AUTHOR Weasenforth, Donald L.
TITLE Prompt Type Effects on Essay Ratings.
PUB DATE May 93
NOTE 78p.
PUB TYPE Reports - Research/Technical (143) --
Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *College Students; *English (Second Language); *Essays; Evaluation Methods; Evaluation Research; Foreign Students; Higher Education; Holistic Evaluation; Influences; *Language Proficiency; Measurement Techniques; Questionnaires; *Rating Scales; Statistical Analysis; *Student Evaluation
IDENTIFIERS *Essay Topics; University of Southern California

ABSTRACT

This study examined 412 college students' essay performance on two prompt types, a traditional prose essay and a type incorporating graphics, modeled on those from English Language Challenge Examination (ELCE) developed for the University of Southern California (USC). The majority of the participants were international students at USC. Each individual wrote one essay based on one of four randomly assigned prompts. Each essay was rated by independent raters using a 10-point criterion-referenced rhetorical scale developed for the ELCE. A subsample of 30 essays was additionally analyzed by means of holistic and quantitative rating scales. Differences between the mean scores on the rhetorical scale were found to be insignificant for all subgroups of participants, including various proficiency levels, academic status, field of study, and various native language groups. Initial analyses of students' responses to two questionnaires and to questions posed during individual interviews indicated variation in familiarity with graphics and some concern about the validity of using graphics as contextual cues with essay prompts to be used for testing. Six appendixes include the essay prompts, student questionnaires, interview questions, prompt evaluation forms, the rhetorical rating scale, and the holistic rating scale. Contains 57 references. (MDM)

Prompt Type Effects on Essay Ratings

by

Donald L. Weasenforth

May 26, 1993

Abstract

This study examined students' essay performance on two prompt types -- a 'traditional' prose type and a type incorporating graphics -- modeled on those from the English Language Challenge Examination (ELCE), the development of which was mandated by the University of Southern California. Each participant in the study wrote one essay based on one of four randomly assigned prompts. Each essay was rated by independent raters using a ten-point criterion-referenced rhetorical scale developed for the ELCE. A subsample of 30 essays was additionally analyzed in two other ways: 1) holistically rated using a criterion-referenced scale of textual abstraction, and 2) quantitatively analyzed through frequency counts of lexicogrammatical features which define textual qualities of informational density and textual abstractness. Differences between mean scores on the rhetorical scale were found to be insignificant for all subgroups of participants, including various proficiency levels, academic status, field of study, and various native language groups. However, initial analyses of students' responses to two questionnaires and to questions posed during individual interviews indicate variation in familiarity with graphics and some concern about the validity of using graphics as contextual cues with essay prompts to be used for testing.

The correlation between mean scores across prompt types for various subgroups of participants was relatively high and approached the parallel-form reliability, indicating insignificant effect on the relative standing of students on the test. These results are supported by a number of omnibus statistical analyses of the data. Prompt type proved insignificant in determining the overall variance in scores for all four types of evaluations of students' essays. Discriminant analyses of all scores and for various subsets of scores indicate that scores can not be accurately distinguished according to prompt type assignment. Finally, a stepwise discriminant analysis did not identify any variables as significant in determining the group membership of scores according to prompt type. A correlational analysis of the four types of evaluations raises questions about the amount of overlap of the two holistic ratings and about the definition of textual abstraction. These results are discussed in relation to the literature on test method characteristics and to practical implications for the ELCE.

Introduction

The English Language Challenge Examination (ELCE) is an achievement test for non-native English speakers that was developed for the University of Southern California. Its purpose is to assess students' abilities to understand formal spoken English, to comprehend written academic texts, and to speak and write in English, within reasonable limits allowable in a standardized testing situation. One component of the ELCE is a 'direct' [1] measure of students' writing abilities. While developing elicitations for this component of the ELCE, there evolved a discussion about the format of the elicitations, with the most interesting focus on the type of contextual cues incorporated in the elicitations.

Given the fact that most of the potential test takers were science / engineering majors, it was suggested that the test would provide a more valid mea sure of academic writing ability if the tasks included interpretation of graphics and the incorporation of information extracted from the graphics in an argumentative discourse. It was then decided that 'graphic' prompts should be developed in addition to the traditional prose prompt. Recent literature, however, has suggested that the use of a prose prompt and a prompt type incorporating graphics would jeopardize the reliability of the writing section of the ELCE. This problem was considered to be important in light of the potential consequences for test takers and the sensitive issues involved in determining students' academic agendas based on limited evaluation of language abilities. This same issue was raised during the deve.opment of the TOEFL's Test of Written English (TWE) (Bridgeman and Carlson 1983, Carlson *et al* . 1985); nevertheless, it was decided that both prompt types would be used on the TWE.[2] However, as a result of concerns expressed about the effect on the reliability of the TWE of using both prompt types (Bridgeman and Carlson 1983, Carlson *et al* . 1985, Greenberg 1986, Raimes 1990), the graphic prompt was eliminated from use in spite of the fact that it continues to appear as an optional elicitation format in the TOEFL literature (Educational Testing Service 1989,

[1] See Bachman's discussion in Bachman (1990) and Bachman and Palmer (forthcoming) of the inappropriate use of 'direct' to refer to the assessment of cognitive abilities, including language abilities. Bachman explains that any measurement of language abilities will necessarily be 'indirect' since the abilities, unlike the performance which manifests them, are not directly observable. To avoid confusion, however, the term 'direct', when used in reference to measurements of writing abilities, will be used in this paper to refer to the use of protocols as opposed to the use of tasks such as cloze or sentence completion.

[2] The graph/chart prompt type was used on the TWE only once for research purposes as part of a comparability study (Raimes 1990).

Raimes 1990). The little existing research which offered insight into the problem included the results of Bridgeman and Carlson's (1983) survey which were used in their determination that the two prompt types, in fact, represented two different types of tasks. Their work, being based on potential raters' impressions of prompts, is more intuitive in nature, requiring empirical evidence to substantiate it. Carlson *et al* . (1985) used protocols collected from the only administration of the TWE in which the graphic prompt type was used to study prompt type effect, but point out the need for replication. This study represents an effort in the direction of teasing out this question through the analyses of protocols which result from the use of the two prompt types.

Several issues involving the comparison of prompt type conditions were of interest in this study. One issue was whether students' essay scores would differ according to the prompt tvpes assigned. If the scores of those students using the prose prompt and the scores of those using the graphic prompt were comparable under the same conditions, it could be suggested that the prompt types do not promote different types of writing as evaluated by the qualitative and quantitative analyses in this study.

Another key issue concerned the correlation in performance under the two prompt type conditions. If the correlation were found to be low in relation to test reliability, it could be inferred that the ranking of students in relation to each other is altered by the variation in assignment of prompt types, in effect changing the character of the test. On the other hand, if the correlations were high in relation to reliability, one could conclude that students' relative standing on the test would not be markedly affected by the choice of prompt types.[3] Parallel-form reliability was assessed to provide a basis for comparison with the correlation between prompt type conditions as discussed in the results section of the paper.

A third issue concerned the reliability of the test under each of the two prompt type conditions. In this regard, information on parallel-form reliability was useful, not only as an aid in

[3] If the ELCE were a norm-referenced test, the correlational data would be the primary evidence used in responding to the question whether the variation in prompt type assignment affects the psychometric integrity of the test. Even if mean scores were affected by prompt type assignment, evidence that the relative standing of students was unaffected by prompt type assignment would, nevertheless, suggest that the essential measurement properties of the test remained unchanged. However, the ELCE writing evaluation scale is criterion-referenced, in that the 'cut-off' score used to determine whether the student should be released from language classes is associated with a particular level of proficiency (i.e., ALI 'high 240' = high advanced level). For this reason, evidence regarding effects on mean performance must be considered along with the correlational data in assessing the merits of using both prompt types on the ELCE writing section.

interpreting the correlation between prompt type conditions, but as valuable evidence in its own right. A substantial reliability difference would be evidence that the use of one prompt type over the other provides greater consistency of measurement. Inter-rater reliabilities would also contribute information related to consistency in the scoring process, and possible differences in the process for essays written under the two conditions.

A fourth issue of interest involved students' reactions to the use of the two prompt types. If the students felt comfortable with the use of the traditional prose prompt, but unfamiliar with graphics and uncomfortable with their use on writing tests, it could be argued that the use of the graphic prompt type decreased the ELCE's face validity from the students' standpoint. To address this issue, students responded to two questionnaires in which they were asked about their familiarity with graphics. Additionally, a subgroup of the students were interviewed to investigate not only their familiarity with graphics, but also their reactions to the incorporation of graphics on writing test elicitations.

A final key issue concerned the correlation of measures of abstraction and prompt type assignment. High correlations between measures of abstraction and prompt types could indicate that prompt types promoted different text types which could partially account for variations in ratings. Textual abstraction could be viewed as a factor of rater reliability latent to prompt type assignment. Several measures of textual abstraction were taken: one qualitative and two quantitative. Biber's (1988) textual analysis model provided a theoretical framework and the analytic methodology for the two quantitative measures as described below. An holistic rating instrument was developed to provide the qualitative measure.

Correlations between the various holistic ratings and textual analysis scores represent another issue with related implicational value. Correlations between the various evaluations (i.e., holistic rhetorical ratings, holistic ratings of abstraction, and the two sets of textual analysis scores) could give some indication of the textual features that are most salient to raters as they score essays. Significant positive correlations between the textual analysis scores and rhetorical holistic ratings, for example, would suggest that the lexicogrammatical features associated with the textual analysis scores play an influential role in determining rhetorical holistic scores. Such information would be valuable given the predominant use of rhetorical holistic evaluations (Perkins 1983).

Although the primary effects of interest were those involving all participants without regard to particular levels of proficiency or other variables, it was also of interest to determine whether the variation in prompt type assignment would have differential effects for subgroups of students defined by: 1) levels of proficiency [4], 2) academic status (i.e., graduate versus undergraduate), 3) field of study (i.e., science / engineering versus other), 4) and native language (i.e., English, Chinese, Korean, Japanese, Indonesian and other). It was hypothesized that particular subgroups of students -- those more proficient in writing English essays (defined by course level), graduates, science / engineering majors, and possibly students from particular language groups -- would be better able to incorporate information from graphics successfully in their writing, possibly leading to better overall performance on the test. To examine the validity of this hypothesis, statistical analyses were completed to look for differential effects according to students' performance on the present essays.

Relevant Literature

One of several significant advances in language testing within the past decade has been the development of a theoretical framework for the description of test characteristics. A systematic description of test facets lays a basis for not only the development of new language tests and the reliable comparison of existing language tests, but also the comparison of non-test related language performance and test performance, and analysis of test method effects on the linguistic performance of test takers. This more systematic approach at describing the characteristics of language tests and their relationship to test takers' performance reflects a change in perspective from seeing tests as a holistic contextual factor to seeing tests as a multifaceted set of contextual factors affecting test takers' choice of linguistic variants.

These productive insights which hold great promise for future development in language testing are the result of a reformulation of the view of tests as artificial linguistic events apart from

[4] Participants provided four types of data which could be interpreted as measures of proficiency in writing English essays. Chronological age, number of years of ESL / EFL training, and number of years of English writing instruction were considered more indirect measures than course level assignment (see Suter 1976, and Purcell and Suter 1980 for further discussion). Since the former variables were insignificant in determining variance in performance across prompt types, they will not appear in this discussion of the study.

naturalistic uses of language to the view that tests should be reflections or samples of authentic speech events. Indeed, the test environment has come to be seen as yet another contextual domain (Bachman 1990, Poole 1990). This reformulation in the way tests are conceptualized parallels developments in the view of context in sociolinguistics as a more interactive and dialogically conceived concept of contextualized language use (Duranti 1985, Goodwin and Duranti, forthcoming).

As linguists showed little interest in the relationship between contextual factors and linguistic competence or performance, the relationship between test methods and test takers' performance received little attention in earlier forms of language testing theory. The major concern in test development was validity, defined largely by the extent to which tests measured a specified propositional domain. There was little concern, however, for the authenticity of tests with regard to the more widely defined target language context, and much less for the reliability of tests as a function of the interaction of test method facets and test taker performance. Indeed, language tests were commonly characterized as static measurement instruments of test takers' competence, the relationship between test and test taker seen as direct and unidirectional (Carroll 1961, Lado 1961).

In current research on language testing, a more dialogic relationship between language user (i.e., test taker) and context (i.e., the test and testing environment) serves as a fundamental theoretical assumption . The performance of any test taker, according to this model, will be determined by not only the test taker's knowledge, linguistic and affective schemata, and the strategies internal to the test taker, but also by test methods and random factors (Bachman 1990, Bachman and Palmer, forthcoming, Oller 1979). Accurate and comprehensive analyses of various test method facets and random factors will allow a more effective partialing out of sources of measurement error so that test scores reflect more clearly the ability testers purport to measure. Consequently scores can be used to make more equitable, accurate decisions.

Two major test facets are of interest in this study of prompt type effects: 1) the response of students to various prompt types, and 2) the response of raters to students' writing. Of particular interest is the first facet, the effect of prompt type on students' written responses. An evaluation of these effects, however, entails an analysis of the writing which takes the form of independent holistic scoring common to standarized tests of writing such as the TWE (Educational

Testing Service 1986, 1989) and ELCE. This study, therefore, investigates prompt type effects as reflected by raters' impressionistic scores and quantitative textual analyses as contrasted with a more psycholinguistic approach to analyze the process of interpreting the prompts and the processes underlying writing ability (Bereiter and Scardamalia 1987, Flower and Hayes 1980).[5]

Responses to various prompt types is in part a function of test taker characteristics. A test taker's performance may vary, for example, with how the student feels at the time of testing, and the test taker's age and cognitive style. These examples represent random factors which are beyond the direct control of testers but which need to be accounted for through analyses of such factors and the application of more sophisticated psychometric measurement techniques.[6] Prompt type effects will be analyzed in this study as partially a function of test taker characteristics. Analyses will produce data for subgroups of students defined by: 1) proficiency level in writing English essays, 2) academic status (i.e., graduate versus undergraduate), 3) field of study (i.e., science / engineering versus other), 4) and native language (i.e., English, Chinese, Korean, Japanese, Indonesian and other). Work by Bereiter and Scardamalia (1987) and Keech (1984) suggests that students more proficient in writing English essays are more capable at incorporating external evidence in their writing. Freedman and Pringle (1980) and Berthoff (1986) suggest that graduate students' maturity will allow for more complex argumentation which will receive favorable evaluations. The survey results of Bridgeman and Carlson (1983) and Horowitz (1986)

[5] This may be an obvious statement, but a very critical observation given the number of factors affecting raters' scores, thus further confounding the analysis of prompt type effects on students' writing. It is, therefore, crucial that reliable ratings be obtained as an indication that sources of error attributable to ratings are reasonably controlled. The quantitative measures are important in that rater error is avoided, providing more objective, albeit limited, analyses of students' texts. Since the goal of the first part of this study is to determine whether prompt type assignment is a statistically significant determinor of rhetorical holistic scores, reliable scoring will suffice to answer the main question addressed by the study. However, to investigate prompt type effects further, a holistic scale of textual abstraction and two quantitative measures of textual abstraction were used. These three measures of textual abstraction were taken to serve as bases for: 1) explanations of statistically significant prompt type effects, and 2) further investigation of textual abstraction as a function of contextual cues, maturity, language proficiency and native language. The literature which presents work addressing these two issues will be briefly reviewed with more attention allotted to the little literature on the specific type of prompt effect of interest in this study.

It should be noted that the two lines of research identified here are seen as complementary. As Bereiter and Scardamalia (1983) have clearly indicated, investigation of the production of written text would be best served by the interface of various levels of research.

[6] A significant limitation of classical methods is the confounding of test taker variables with test method characteristics. Item Response Theory (IRT) models allow for independent measurement of linguistic abilities and factors affecting actual performance. A three parameter IRT model, for instance, controls for guessing (Hambleton, Swaminathan and Rogers 1991).

reflect a differential in usage of graphics by science / engineering students versus students of other disciplines. This difference implies that science / engineering students may be able to extract information from graphics and perhaps integrate it into their writing more adeptly than can other students. Finally, Bridgeman and Carlson (1983) point out that students from various cultures may differ in their familiarity with and use of graphics, a factor which could directly affect the overall quality of responses to graphic prompts. This concern is also reflected in studies of contrastive rhetoric. Whether students use information incorporated in graphics, the type of information used, the extent to which it is used, and the manner in which it is incorporated in the discourse of their own writing are all partially determined by cultural constraints on discourse production (Duranti 1985, Ferris 1989, Kaplan 1966).

Students' performance on tests is also a function of test method characteristics, test methods including the format of instructions, salience of test sections, and the amount of contextualized information in test items, for instance. Test method facets can be directly manipulated by test developers in order to minimize measurement error which results, in part, from uncontrolled test taker characteristics. Test method facets which have been proposed as significant determiners of test performance include topic assignment (Freedman 1977, Freedman 1979, Hoetker 1982, Ruth and Murphy 1988), specification of rhetorical constraints (Flower and Hayes 1977, Flower and Hayes 1980, Odell 1981), specificity of instructions (Greenberg 1982), prompt length (Brossell 1983, Brossell and Ash 1984), wording of prompts (Greenberg 1982, Ruth and Murphy 1988), chronological presentation of a series of prompts (Hayward 1989, 1990), the cognitive complexity represented by the task (Bereiter and Scardamalia 1987, Bridgeman and Carlson 1983, Caccamise 1987, Keech 1984; Quellmalz, Capell and Chou 1982; Tetroe as quoted in Bereiter and Scardamalia 1987), situational constraints (Nelson 1990) and mode of production (Moustafa, 1987).

The test method facet which is of primary interest in this study is the contextual cues offered in essay prompts. Specifically, a traditional prose prompt and a prompt incorporating graphics will be compared. These two prompt types were used in this study due to the fact that they were identified as most representative of prompts used for academic essay writing (Bridgeman and Carlson 1983, Horowitz 1986), the type of writing measured by the ELCE. Opposition to the use of both prompt types on the TWE prompted Carlson *et al* . (1985) to investigate prompt type effects as part of a broader study of the relation between admission test scores and writing proficiency. They used four types of evaluations -- rhetoric al holistic rating, discourse and sentence level ratings, and computer scoring[7] -- of data collected from the only administration of the graphic prompt on the TWE. Contrary to implications drawn from Bridgeman and Carlson's (1983) work, Carlson *et al*. found no statistical evidence that would indicate a significant difference between the two prompt types. They attribute this finding to the fact that they developed both prompt types with the intention of reflecting academic tasks.[8] They conclude that further research is needed before prompt type differences can be dismissed as relatively insignificant and recommend that their results be replicated for individual programs.

> The results of this study could be interpreted to suggest that performance on one writing assignment provided valid and reliable information regarding performance on the other tasks; with new topics, a different ... population, and under slightly different testing conditions, however, this finding would need to be demonstrated (Carlson *et al*. 1985: 81).

These findings and conclusions are supported by evidence from Hale's (1991) study of the effect of the amount of time on TWE essays. As a subcomponent of his study, Hale included a special condition using graphic prompts.[9] Although Hale finds no statistically significant

[7] The discourse rating represented an evaluation of the organization of textual material in assuring coherence; the sentence level evaluation focused on grammatical and mechanical correctness. Writer's Workbench software was used to analyze a subsample of essays for total length, average sentence length and the correctness of grammatical forms such as subject-verb agreement.

[8] The work of Bridgeman and Carlson (1983) and Carlson et al. (1985) identify their focus of investigation as test reliability, raising a more fundamental question about the components of a test's usefulness as defined by Bachman (1990, Bachman and Palmer, forthcoming). Reliability is but one factor, along with validity, authenticity, impact and practicality, which determine a test's overall usefulness. The issue addressed in these two studies and in the present study could be more accurately defined as involving not only reliability, but also validity and authenticity. The earlier work more directly addresses the issues of the TWE's authenticity and construct validity. That is, the use of graphic prompts was found to be a more authentic academic task for most disciplines represented in the study than was the task of comparison / contrast for example. Use of the graphic prompt is also assumed to lead to more valid interpretation of scores. The issue of construct validity is not explicitly addressed although it is a very important issue in this line of research. There are two important considerations in this regard: 1) whether the interpretation of graphics should be considered part of the construct of 'writing ability', the ability purportedly tested by the TWE, and 2) to what extent should the validity of the TWE be sacrificed to provide for more authentic writing tasks. The reliability issue addressed is more specifically that of inter-rater reliability.

[9] It should be noted that Hale uses a different type of graphic than those used by Bridgeman and Carison (1983), Carlson *et al*. (1985), and those used in this study. Whereas the graphics used by Hale incorporate information largely in linguistic form, the graphics incorporated in the prompts used in the

difference in inter-rater correlations, student responses to questionnaires reflect variation in students' reaction to the two prompt types. While the 'academic' students indicated that 30 minutes was sufficient for writing an essay based on the graphic prompts, the 'intensive' students saw this amount of time as insufficient. Both groups of students were, however, in agreement that 30 minutes were sufficient to write an essay based on the prose prompts. This finding suggests that students' reactions to prompt types is partially a function of language proficiency level.

Before continuing with summaries of other areas of the literature which have informed this study, a note about the general lack of research in this area is due. The inconsistency of and general lack of research in this area can be traced to several sources, some of which are restrictions on research in testing in general. One reason for the lack of literature is ETS's decision to drop the graphic prompt type from the TWE. The research generated by the controversy over the decision to use the prompt type has become incidental, as in Hale's (1991) work, as more pressing issues have been addressed. It is also true that the nature of the problem is very complex, involving the difficulty of describing the control of tasks through reading. Research in prompt effects is, thus, splintered due to the multifaceted nature of the issue.

Other reasons are common to research in testing in general. The collection and analysis of data is very time consuming and very costly. Unfortunately, it is a task which is repeated out of necessity due to the unavailability of possibly adequate data which may have already been collected. The unavailability is in part due to the reluctance of testing agencies and educational districts to provide data to outside researchers. Another problem is that research in testing has been largely carried out by individuals who work to fulfill separate agendas (Skehan 1990). Perhaps one of the most formidable barriers facing researchers in this area has been the lack of a theoretical framework as a basis for the analysis of prompt type effects. This gap is now being filled by work on test method analyses (Bachman 1990, Bachman and Palmer, forthcoming).

Another key facet which plays a crucial role in this study, although not the primary focus thereof, is the rating of students' essays. Various aspects of this facet have received attention in the literature (Freedman 1977, Freedman 1979, Perkins 1983, Winters 1979). As with student responses, many aspects (e.g., topic bias, internal lack of consistency, and shifting standards from

other three studies include bar and line graphs, and pie charts. This should be considered a factor in responses to the prompts (Bachman 1990, Bachman and Palmer, forthcoming).

one paper to another) of this facet can not be directly controlled although training and reliability measures may be used to attenuate the effects of these factors. One such aspect of this facet which has received a great deal of attention in the literature are the characteristics of protocols which raters focus on in the process of scoring. Diederich *et al*.'s (1961) work is particularly noteworthy in that it identifies a number of factors that affect essay raters. Freedman (1977, 1979) found evidence of a hierarchical ordering of textual features which determine the scores assigned by raters. Other researchers have identified specific factors as being generally influential in determining scores. Hake and Williams (Hake and Williams 1981, Williams 1979), for instance, have suggested and presented evidence to the effect that the organization of propositional content in protocols is an important source of rating variance. Their research suggests that the complexity or abstract nature of texts is a significant factor in determining rating variance. More specifically, their results suggest that texts with more abstract discourse generally receive higher ratings.[10] This study also investigates textual abstraction, defined in terms of the complexity of argumentation, as a source of variance in raters' scores. Before discussing the theoretical basis for the measurement of abstraction used in this study, however, a brief description of an earlier attempt to measure levels of textual abstraction may be instructive.

Freedman and Pringle (1980) defined abstraction in terms of the levels of generalization of propositions presented in students' essays. Their main interest was the correlation of observed uses of abstraction (defined as higher levels of generalization) with maturity levels of high-school and college students, maturity levels being operationally defined as the number of years of formal education completed. They hypothesized that more mature students would produce writing in which they would present original generalizations encompassing arguments and supporting evidence. More mature students, for example, would relate arguments and discrete pieces of evidence to a general framework, whereas less mature students would present disparate pieces of information without reference to a coherent theoretical framework. The results of their study were

[10] Hake and Williams (1981) look specifically at nominalization and the organization of 'given' and 'new' information within a text, claiming that increased use of nominalization and the unsystematic presentation of 'given' and 'new' information leads to more abstract texts which is encouraged and rewarded by composition instructors. Williams (1985) also suggests that the inconsistent assignment of agent thematic role to subjects of sentences can likewise lead to more abstract texts.

inconclusive, partially as a result of operationalization difficuities.[11] Nonetheless, their work has served as a useful guideline for the present study.

Of particular interest to the analysis of abstraction levels was theoretical work in the structure of argumentation (Tirkonnen-Condit 1985, Toulmin 1958) which served as a basis for the construction of an holistic instrument used in the evaluation of argumentation presented in the protocols. Toulmin provides a structural analysis of argumentation as a cognitive process of problem-solving which has been successfully applied to the analysis of L_2 discourse (Connor 1987, Ferris 1990) as well as L_1 discourse (Tirkonnen-Condit 1985). According to this theory of argumentation structure, adequate argumentation consists of components of 'situation' (i.e., discussion of the nature and history of the undesired condition and / or proposed changes), 'problem development' (i.e., explication of the undesired conditions), 'solution' (i.e., presentation of proposed changes), and 'evaluation' (i.e., discussion of constraints on the proposed changes). The main activities of building argumentation include formulating a proposition, collecting and marshaling data, analyzing data and problems, and formulating a solution. The holistic rating scale of abstraction incorporates these ideas along with elements from Delia, Kline and Burleson's (1979) scale of audience address.

The line of research which has provided this study with a framework for the quantitative analyses of abstraction is that of Biber's (1988) computer-aided textual analysis. Biber's Multifeature / Mulitdimension (MF / MD) model provides a framework through which texts can be described along a number of dimensions according to the number and type of lexicogrammatical features that cluster within a text. The protocols used in this study were analyzed along two dimensions defined by Biber to provide quantative analyses of the texts' abstractness. Biber's Dimension 1 represented by a continuum defined by the rubrics 'involved' and 'informational' describes the information density of a text, one measure of textual abstraction assumed by this study. An analysis of features defining the dimension (i.e., Biber's Dimension 5) of textual

[11] Two notes should be made with regard to Freedman and Pringle's (1980) study. First, it was not intended to inform any testing model although this seems theoretically and practically feasible. Within Bachman's (1990) framework, this issue would be considered a facet of 'expected response.' Second, the operationalization of the constructs of abstraction and maturity is ambiguous, a key factor in determining the inconclusive results of their study. Although they claim to analyze abstraction "as evidenced in [students'] writing" (Freedman and Pringle 1980: 318), they divorce the cognitive abilities associated with abstraction from discourse, the vehicle for the manifestation of abstraction. This problem is discussed in more detail in Berthoff (1986).

'abstractness' - 'nonabstractness' will provide a second measure of the abstract nature of texts. Abstractness defined along this continuum is typified by syntactic marking of "complex logical relationships that characterize texts of a more technical, formal nature" (Biber 1988: 112). Correlations between these two analyses and prompt types will be analyzed to determine whether the individual prompt types promote the production of a particular type of discourse.

Method

Subjects

The data used in this study were provided by 412 students, 178 students in Freshman Writing Program (FWP) courses and 234 in American Language Institute (ALI) courses at the University of Southern California.[12] Thirty-two of the FWP students were native English speakers. The remaining 380 students were international USC students, both graduate and undergraduate, enrolled during the Fall Semester of the 1991-1992 academic year. Undergraduate students comprised 65% of the 412 participants but only 38% of the ALI subgroup.

Proficiency levels for the international students ranged from intermediate university ESL level to second semester freshman writing level. The native speakers were all completing their first semester of freshman writing courses.

Students were asked to complete two questionnaires and write one essay as part of course requirements. Students were assured of the "practice value" of writing the essay. Copies of essays were returned to students later, along with a summary of the ratings and raters' comments for each student who participated in the study. Participants in the interviews were recruited in the classes that participated in the study and were offered $10 as an incentive for participation.

Native languages of the international students included several major dialects of Chinese (with Chinese speakers comprising 48% of the total sample), Korean (15%), Japanese (8%),

[12] A total of 421 students actually participated in this study. All 421 observations were processed in the analyses of inter-rater reliability. For all other statistical analyses, nine observations were dropped due to uncharacteristic performance of the participants as determined by a comparison of midterm/final grades and scores assigned for this study. Data associated with USC employees were part of the data dropped since these participants represent a population different from that of university students which was of interest in this study.

Indonesian (7%), and 17 other language groups with fewer than 5% in each. Native countries represented were Taiwan (30% of the sample), Republic of Korea (16%), the People's Republic of China (13%), Japan (8%), Indonesia (7%), Hong Kong (6%), and 28 other countries with fewer than 5% from each.

Materials

Prompts. Four essay prompts were used in the study: two "traditional" prose prompts and two "graphic" prompts. The traditional prompt consisted of a prose description of the topic and instructions, including rhetorical constraints. The graphic prompt consisted of the same prose description with additional instructions related to the use of the graphics and the accompanying graphics. Each graphic prompt incorporated three graphics, a combination of bar graphs, line graphs and pie charts.

The two topics -- "homelessness" and "fresh water supply" -- were used to reflect two major fields of study represented by the student population. The "water" topic reflects interests in science and technology while the "homelessness" topic relates to general social science interests. Topics and prompt types were both randomly assigned to participants. The four prompts are presented in Appendix 1.

Essays. Two lined sheets of paper were provided to each student for writing essays. Additional sheets were provided as needed. Students were asked to indicate both the student number assigned for the study and the prompt number in spaces provided in the upper right hand corner of the first lined sheet. Prompts appeared on separate sheets which were randomly distributed with the first questionnaire.

Questionnaires. Participants completed two questionnaires, one immediately before and one immediately after writing the essay. The first questionnaire requested biographical information and responses to questions regarding interpretations of the prompt. The second questionnaire requested responses to questions regarding interpretations of the prompt and their effect on the participant's writing. The questionnaires are presented in Appendix 2.

Interviews. Participants were encouraged to take part in two individual interviews, one at least 24 hours prior to writing the essay and one within 24 hours after writing. Questions used in the first interview elicited responses regarding interpretations of the particular prompt randomly assigned to the participant and the usefulness of the prompt in writing. Participants were not informed that they would be using the prompt for writing in class. The second set of interview questions investigated the possible changes in interpretations of the assigned prompt and the actual use of the prompt in writing. Both interviews of participants assigned graphic prompts also incorporated questions about the participants' familiarity with graphics as well as the usefulness of the graphics in writing the essay. A total of 20 students participated in at least one interview with a total of 26 interviews completed. Both sets of questions are presented in Appendix 3.

Prompt evaluations. Evaluations of the prompts were requested of experienced composition instructors not only as a means of editing but also to provide information useful in analyzing participants' interpretations and actual use of the prompts. These evaluations provided the third leg of the "triangulation" design of the study, supplementing information gathered in the questionnaires and interviews and in the analyses of the writing samples. The prompt evaluation form is presented in Appendix 4.

Traditional rhetorical holistic rating instrument. The rhetorical holistic rating instrument used in this study was developed specifically for the ELCE. It consists of a ten point, criterion-referenced scale with descriptors for "organization," "content / development," "grammar / vocabulary," and "mechanics."[13] A copy of the original version of the instrument used in this study is presented in Appendix 5.

[13] Since its use in this study, this instrument has been significantly abbreviated so that ratings can be completed more quickly, but the original ten point scale and most of the original descriptors have been retained. The revised instrument does, however, incorporate a reader-oriented descriptor of "general intelligibility." Since most of the basic characteristics of the scale have been retained along with the original theoretical view of essay evaluation, it is assumed that ratings on the two versions of the scale would be highly correlated, although this obviously should be empirically substantiated.

Holistic rating instrument to evaluate levels of abstraction. A second holistic rating instrument was developed to provide an analysis of the complexity of the argumentation used in writing samples. This instrument also consists of a ten point, criterion-referenced scale. The descriptors are: "proposition," "data," "analysis," "argument structure," and "audience awareness." A copy of this instrument is found in Appendix 6.

Research Design

The treatment groups in the study are shown in Table 1; entries in the table are the numbers of students per subgroup. As shown in the headings in Table 1, some students were given traditional prompts, and others, graphic prompts. Within those general groups, some students were given the "homelessness" topic while others were given the "water" topic.[14]

Control group. Native English speakers in the FWP served as a control group (first row of Table 1). This group wrote one essay using either the traditional or graphic prompt type, but only the "homelessness" topic was used due to the incompatibility of the "water" topic with the FWP 101 curriculum.

Experimental group. In the experimental groups (rows 2 - 8 of Table 1) each student wrote one essay using one of the four prompts developed for the study. Note that, in an ideal design, the

[14] The original design for the study included repeated measures so that each student would write two essays, one using a "traditional" prompt and another on the same topic but using a "graphic" prompt. Since this design proved impractical for both FWP and ALI programs due to the amount of class time this would have required, a single measure was used with analyses based on an assumed approximate equivalence of abilities within proficiency levels. Proficiency levels were defined for this study by performance on the essay written for the study. This measure correlated well with midterm / final grades collected for FWP 101 (i.e., first semester freshman writing course) NSs of English, FWP 101 NNSs, ALI 221 (i.e., intermediate ESL course) and ALI 220 (i.e., intermediate ESL course for graduate science majors) levels. The correlations were not consistent for FWP 111 (i.e., second semester freshman writing course for NSs) and ALI 240 (i.e., advanced ESL course) levels, however. Students of these levels were assigned proficiency levels based on rank ordering of the scores that were assigned to the essay written for this study.

numbers per group would be equal, so that the different prompts would completely balance each other. However, use of least squares analyses of variance effectively ensured that the different orders contributed equally in computing effects of the factors under investigation.

Table 1
Research Design

	Traditional Prompts		Graphic Prompts	
	Water topic	Homelessness topic	Water topic	Homelessness topic
Control Group[15]				
FWP 101 Students	N = 0	N= 17	N = 0	N= 15
Experimental Subgroups by Proficiency Level				
FWP 101 NNS	N = 0	N = 8	N = 0	N = 9
FWP 111 / 112 / ALI 262 - Group 1	N = 17	N = 16	N = 22	N = 14
FWP 111 / 112 / ALI 262 - Group 2	N = 19	N = 24	N = 21	N = 21
ALI 240 - Group 1	N = 20	N = 22	N = 26	N = 20
ALI 240 - Group 2	N = 22	N = 13	N = 21	N = 14
ALI 221	N = 5	N = 7	N = 7	N = 8
ALI 220	N = 8	N = 8	N = 4	N = 5

Procedure

The students were told that they would be asked to write in response to a prompt which was to be randomly assigned to them and to complete two questionnaires, one before and one after writing. The students were informed that the score assigned by raters would not affect their course grades, but that it was, nonetheless, important that they do their best, as the results of the study would, in part, determine how prospective USC international students would be tested in the

[15] Note that for FWP 101 students, represented in rows 1 and 2 of Tables 1 and 4, the water topic was not used. This topic was considered inappropriate for that particular population.

future. Students were also informed that copies of their writing would be returned to their instructors to be used for whatever purpose the instructor thought was appropriate.

Students were then given a copy of questionnaire 1 and randomly assigned both an ID number and prompt (i.e., both topic and prompt type were randomly assigned). After all students completed the questionnaire, the questionnaire form was collected and the lined paper was distributed. All students were allowed exactly 35 minutes to plan and produce a writing sample and were reminded of the time limit when 30 minutes had passed. After writing, students were asked to complete questionnaire 2 and were dismissed as they finished.

Essay Scoring

Each essay represented in this report was scored holistically by two sets of experienced independent raters on two ten-point, criterion-referenced scales. Two different sets of raters were used so as not to confound the evaluations with effects associated with multiple readings / ratings by the same raters. A total of five raters were involved, with the researcher serving as one of the raters for both sets of raters. Problems associated with confounding the two ratings by using a common rater for both sets of raters, however, were minimized due to the fact that two months period separated the two types of ratings. Each rater received a copy of each of the essay prompts and a copy of the appropriate rating scale. Each rating session was preceded by a norming session during which all raters discussed the format and theory underpinning the rating scale, then scored 5 - 10 essays, after which scores were compared and discussed in relation to the rating scale.

The first set of raters completed the traditional rhetorical holistic evaluation of all 412 original essays during one weekend. Raters scored essays in batches of approximately 50 essays each, alternating between the "water" and "homelessness" topics. In addition to the norming session at the beginning of the day's rating session, there was a norming session before rating each batch of papers, using approximately 3 - 4 essays drawn from the batch to be rated. Scores were recorded on separate score sheets. They were collated at the end of each day as a measure against inaccuracies in recording and to provide a basis for a rough estimate of rater reliability.

The second set of raters evaluated a subset of 30 essays drawn from the original 412 essays through stratified random sampling. The second rating was based on a holistic evaluation of the complexity of argumentation, using a holistic instrument focused on argumentation structure, the cogency of arguments and audience address. This rating was completed during a 48 hour period at the discretion of individual raters after the completion of a two hour norming session. During the norming session, raters were cautioned not to apply criteria generally associated with rhetorical scales, that is, they were asked not to base scores on evaluation of sentence level errors, such as grammar errors, punctuation errors, or errors in vocabulary choice.[16] The rather lengthy norming session was due to the novelty of the rating scale. Each rater had extensive experience in rating essays on rhetorical scales, but there was little apparent commonality between this scale and rhetorical scales, so that their experience with rhetorical scales was largely irrelevant in applying the novel scale. The novelty of the scale prompted a number of questions about the theoretical basis for the evaluation and resulted in some difficulties in applying the scale adequately to ensure acceptable rater reliability. Nonetheless, the reliability statistics for this rating presented in Table 5b suggest that the norming session was relatively successful. The same procedure for recording and collating raters' scores was followed as had been for the rhetorical rating.

For the rhetorical ratings, essays which were assigned scores differing by three points or more were rerated independently. Although raters were aware of the rerating procedure, they were given no information regarding the original disparate scores. If scores still differed by more than three points for any rerated essay, it was decided that the essay would not be used for further analysis and that the scores would be excluded from the data base. It was not necessary to exclude any of these scores from the data base, since the scores for all rerated essays fell within a less than three point range. No essays were rerated for the rating of abstraction due to logistical difficulties. All initial scores for this rating were included in the data base.

Scoring for every essay was "blind" with respect to all important factors: a) the assigned prompt type (i.e., prose vs. graphic), b) the student's native language, c) the student's academic status, and d) the student's proficiency level. To ensure blind scoring, the indication of prompt type on the lined paper was coded so that this information would not be interpretable. The

[16] The second set of raters was asked to ignore, to the best of their ability, errors in grammar, punctuation, paragraphing, and vocabulary choice. It was believed that this was necessary in order to distinguish the two holistic ratings in spite of an awareness that a strict distinction between discourse, a defining element of abstraction, and grammar, punctuation, paragraphing and vocabulary choice is artificial (Celce-Murcia 1990). The decision made to distinguish discourse from these other linguistic categories is consonant with Bachman's (1990) model of linguistic competence.

student's identification number was the only other piece of identifying information on the lined paper, but this information did not affect the "blind" reading since numbers were randomly assigned to all students. Furthermore, for each topic, all essays had been shuffled before scoring to ensure that essays were not grouped according to course levels.

MF/MD Analyses

Two quantitative measures of textual complexity were used in addition to the holistic evaluations, one a measure of informational density and the other a measure of syntactic abstraction. Both types of measures rely on frequency counts of lexicogrammatical items which cluster to define texts along a number of dimensions. One of these dimensions is represented by a continuum along which texts are defined according to informational density. The poles of this continuum are labeled 'Informational Production' and 'Involved Production.' The second dimension used in this study represents the 'Abstract versus Non-Abstract Information' nature of texts.

These analyses required frequency counts of each grammatical and lexical features associated with each dimension.[17] The original identification of dimensions was completed through Biber's application of exploratory factor analysis with a large corpus of both written and spoken texts. The factor analysis procedure identified clusters of lexicogrammatical features which constituted manifestations of textual dimensions. The definition of dimensions was based on information from previous research on the functional and distributional characteristics of the various features. The 'Abstract-Nonabstract' dimension, for instance, was defined by the clustering of features such as agentless passives and past participial WHIZ deletions which highlight the relationship between the patient of a statement and the discourse topic, and are frequently found in abstract, technical discourse such as official government documents.

[17] The 'Informational - Involved' dimension was defined by features such as private verbs, THAT deletion, contractions, and present tense verbs which have positive inter-factor correlations, as well as by features such as nouns, word length and prepositions which have negative correlations. There are no features associated with the 'Abstract-Nonabstract' dimension with negative inter-factor correlations. The features which define this dimension are conjuncts, agentless passives, past participial clauses, BY-passives, past participial WHIZ deletions (i.e., participial clauses functioning as reduced relatives), and other adverbial subordinators.

Frequency counts were first normalized for a text length of 1,000 words so that the salience of features could be compared across texts of varying lengths.[18] It would be meaningless, for example, to compare the 50 occurrences of private verbs in a text of 500 words with the 50 instances of the same feature in a text of 1000 words, since private verbs would not be as numerically salient in the longer text. Normalization, then, compensates for variations in text length, allowing for a valid comparison of numerical salience of a feature from one text to another.

Normalized values were then standardized to allow for comparison of the salience of one feature with another.[19] Before standardization, the frequencies for various features represent different scales with differing means and standard deviations which does not allow for a valid comparison of different features. Without the standardization of frequencies, it would not be possible to conclude that the occurrence of 50 nouns and 100 prepositions indicates that prepositions are numerically twice as salient in a particular text. Standardization converts frequencies of all features to a common scale with a mean of 0.0 and a standard deviation of 1.0, allowing for the validity of such statements.

The scores which represent the 'informational' and 'abstract' nature of the texts were derived by totaling the standardized, normalized frequency counts for features on each dimension. The 'informational' score was derived by totaling the manipulated frequency counts for each feature with a positive inter-factor correlation and subtracting the manipulated scores for the features with negative correlations. Since there were no features associated with the 'Abstract-Nonabstract' dimension which had negative inter-factor correlations, the 'abstract' score was derived by totaling the manipulated scores for all features.

Statistical Analyses

Several types of multivariate analyses were performed on the data to identify sources of variance as a factor of prompt type effect and to determine whether essay scores could be distinguished based on prompt type assignment. A multiple linear regression and multiple

[18] Normalized frequency counts were computed as follows:
Normalized frequency count = Raw frequency count / Number of words in text x 1000.

[19] The standardized score for each feature was computed as a standard z-score:
Standardized score = Normalized frequency count - Mean / Standard Deviation.

correlations matrix were run under the Statistical Analysis System (SAS) procedure, general linear model (GLM) (Statistical Analysis System 1989), to provide two types of information: 1) a statistical measure of the model's reliability, and 2) identification of the factors which were significant in accountnig for variations in scores. Discrimant analyses were run on Statistical Package for the Social Sciences (SPSS) (1990) to determine if the scores could be classified according to prompt type assignment at a statistically significant level of accuracy. Finally, a stepwise discriminant analysis was completed using SAS to identify variables most effective in distinguishing scores according to prompt type assignment as a function of the amount of variance in the scores for which they accounted for.

Results

Inter-Rater Reliability

Before considering the results of the principal analyses, it is useful to examine data relating to inter-rater reliability. Three statistics were computed: a) the discrepancy rate, which is the percentage of papers for which the three ratings differed by three or more points[20], b) the average of correlations[21] between scores given by the three readers, and c) the coefficient alpha reliability.[22]

[20] The discrepancy rate quoted should be distinguished from the similar descriptive statistic provided by Educational Testing Service. Whereas the ETS statistic is equal to one-half times the percentage of papers for which the ratings differ by two or more points on a six point scale, the discrepancy rate used in this study is simply the percentage of paper for which the initial scores differed by three or more points on a ten point scale. Both sets of data are nonpsychometric statistics which give a more general indication of inter-rater reliability.

[21] Pearson product-moment coefficients were calculated as a statistical basis for correlation analyses. The use of the Pearson *r* versus a non-parametric analysis was warranted by the fact that the distributions of rhetorical scores for the subsets of data analyzed were near normal. This provided a basis for the validity of the other parametric procedures including the regression analysis and discriminant analyses. The statistical analyses of the other evaluations (i.e., the holistic evaluation of abstraction and the two MF / MD analyses) are very tentative in light of the non-normal distribution of these data for various subgroups as reflected in the means and standard deviations recorded in Tables 6b, 7a, and 7b. A substantial basis for such statistical analyses will be provided as work on the study procedes.

[22] The standardized alpha coefficient, Cronbach's alpha, was computed as follows:

$$\text{Alpha} = 1.5 (1 - [S^2_1 + S^2_2 + S^2_3] / S^2_t$$

where S^2_1 . S^2_2 and S^2_3 refer to the rating variances of the three raters, and S^2_t refers to the variance of the sum of the ratings.

The inter-rater reliability results of the rhetorical scoring are shown in Table 2a; results of the evaluation of abstraction levels in Table 2b. Data are first presented for all essays, within traditional and graphic prompt types for topics combined and then for each individual topic.

Generally, the figures are acceptable for a standardized test like the ELCE, except the measures for the "water" topic with the graphic prompt which represent the only relatively low measures of reliability. The consistently higher discrepancy rate and lower inter-rater correlation and alpha reliability measures for the "water" topic-graphic prompt combination are apparently partially due to the differential in familiarity with the topic among raters. With the use of "water" essays during norming sessions for the rhetorical rating it was apparent that the scores of one of the raters varied significantly from the other two scores, admittedly due to lack of familiarity with the topic. Nevertheless, the inter-rater reliability measures still show this topic in combination with the traditional prompt type and other topic-prompt combinations to provide adequate levels of inter-rater reliability. It appears, however, that the lower reliabilities for the "water" topic-graphic prompt combination may be more a result of the format or content of the prompt than rater knowledge differences, since the "water" topic-prose prompt combination yielded a high inter-rater correlation (.91) with the ratings of abstraction. A future detailed facet analysis of the prompts and expected responses may yield a satisfactory explanation for this discrepancy. No apparent explanation is available at this time.

Aside from the "water" topic-graphic prompt combination, comparisons across prompt types show little difference between them with regard to raters' scores. Inter-rater correlations associated with traditional prompts (.78 for both analyses) and graphic prompts (.74 for the rhetorical analysis and .80 for the analysis of abstraction) do not represent significant variance in raters' impressionistic evaluations of essays. Apparently, then, the use of graphic prompts did not significantly affect scoring reliability. Instead, the measures of inter-rater reliability indicate that topic was more influential in determining variation in raters' scores than was prompt type. This finding is consistent with results from earlier regression analyses of the data which identified topic as a relatively significant factor in determining the variance associated with scores.

An analysis of inter-rater correlations within topics yields evidence that supports conclusions drawn from correlations across prompt types without regard to topic. Looking first at data from the rhetorical rating (Table 2a), the average correlations for the "homelessness" topic across prompt types are very similar (.80 for the traditional prompt and .84 for the graphic prompt). The difference between average correlations for the "water" topic across prompt types nears significance (.74 for the traditional prompt and .64 for the graphic prompt).[23]

However, data produced by the analysis of abstraction levels (Table 2b) allows for some interesting observations which do not support conclusions drawn from analyses of the correlations between the rhetorical ratings. Although the correlations for prompt types without regard to topic are similar (.78 for the traditional prompt type and .80 for the graphic prompt type), comparison of average correlations within topics indicate significant differences. The average correlations for the "water" topic are .91 (for the traditional prompt type) and .66 (for the graphic prompt type). Similarly, the average inter-rater correlations for the "homelessness" topic are .69 (for the traditional prompt type) and .84 (for the graphic prompt type). However, the fact that the differences in the within-topic correlations are not unidirectional indicates that prompt type is not the only factor accounting for the difference. The high correlations (.84) for the "homelessness" topic-graphic prompt type combination and (.91) for the "water," compared to the other two relatively low correlations (.69 and .66) imply that prompt type and topic are at least two factors which determine the reliability of raters' scores. These results may be the result of a relatively small sample size.

It would appear that the reliability of content-focussed evaluations of protocols is significantly affected by prompt type (perhaps as a cofactor of topic) unlike those of rhetorical evaluations. This may not be surprising since content is only one of a number of equally weighted descriptors on the rhetorical rating instrument. Assuming that the raters reliably employed the

[23] Based on the fact that passing the ELCE is one of several ways to avoid the requirement to complete language classes, and the possible release of students during reevaluation, this difference is considered insignificant. Prospective ESL students may bypass language course requirements by performing adequately on the Incoming Student Examination (ISE) or by obtaining a waiver from his/her academic department. Students who are placed into ALI classes based on their performance on the ISE may still be exempted from course requirements during the first week of classes. All ALI students are reevaluated during the first week by the instructors of the classes for which they have registered. Students who perform adequately on the reevaluation examinations may avoid language class requirements. Given the fact that repeated evaluation helps insure that an adequately proficient student will not be required to take unnecessary language courses, the relatively low inter-rater reliability correlation (.64) associated with the "water" topic - graphic prompt combination could be viewed as acceptable. A student, for example, who might be required to take a writing course due to low scores which resulted from relatively low reliability between / among raters would still have several ways of avoiding the requirement. See Bachman and Palmer (forthcoming) for a discussion of the compensatory relationship between reliability and impact as factors of the usefulness of a test.

rhetorical scale in rating the essays, the other descriptors may have served to anchor the ratings, resulting in less variance in scores across raters.

Since this type of instrument is not currently used in rating protocols, little concern needs to be generated over the effect of prompt type on content-focussed evaluations. These observations do, however, raise interesting questions regarding the role of content evaluation in essay ratings. Further analyses may also yield engaging insights into the relation between the provision of contextual cues and the production of written discourse.

Another important issue in the analysis of inter-rater reliabilities using the evaluation of abstraction levels concerns the novelty of the instrument. The rhetorical instrument is commonly used by composition instructors and essay raters and, therefore, is more familiar to raters than is the instrument recently developed for this study. The relatively high inter-rater correlations imply that the instrument developed to assess levels of abstraction is generally reliable, but the lack of familiarity with the new instrument may be a factor in accounting for the variance in scores. The validity of these conclusions will be tested once a larger sample of essays has been rated using the scale of abstraction.

Table 2a
Discrepancy Rate, Inter-rater Correlation and Alpha Reliability for Major Samples of Papers:
Traditional Rhetorical Holistic Evaluation

	Discrepancy rate	Inter-rater correlations (average)	Alpha Reliability
All essays (both prompts)	4%	.76	.90
Prose prompts			
All essays	2%	.78	.91
Water topic	3%	.74	.89
Homelessness topic	1%	.80	.92
Graph prompts			
All essays	[illegible]%	.74	.90
Water topic	7%	.64	.84
Homelessness topic	4%	.84	.93

Table 2b
Discrepancy Rate, Inter-rater Correlation and Alpha Reliability for Major Samples of Papers:
Holistic Evaluation of Abstraction Levels

	Discrepancy rate	Inter-rater correlation (average)	Alpha Reliability
All essays (both prompts)	47%	.79	.92
Prose prompts			
All essays	60%	.78	.91
Water topic	33%	.91	.97
Homelessness topic	67%	.69	.86
Graph prompts			
All essays	33%	.80	.92
Water topic	40%	.66	.77
Homelessness topic	30%	.84	.93

Parallel form reliability

Parallel form reliability was calculated as a measure of the prompt types' usefulness in promoting consistent measures of students' writing abilities. T-tests were run to determine whether a significant statistical difference existed between the means of rhetorical scores across topics and across prompt types. Table 3 below presents parallel form reliability data for both topics (rows 1 - 2) and prompt types (rows 3 - 4). None of the figures indicate a significant difference in scores, neither across topics, nor across prompt type. The 2-tail probability for between prompt type analysis for the "homelessness" topic is .97, indicating virtually no statistical difference in scores assigned to essays written about this topic as determined by prompt type. The two-tail probability for a comparison of scores across prompt types for the "water" topic is .79, indicating an insignificant statistical difference between scores assigned to essays based on the "water"-prose prompt and scores assigned to essays based on the "water"-graphic prompt. These results suggest that both the prose and graphic prompts, in combination with either topic (further supported by the 2-tail probabilities in a comparison of results across topics, .92 and .71), allow for consistent rating. That is, ratings are not significantly skewed due to the use of both prompt types. The consistent relatively high inter-rater reliabilities are consistent with these results.

Table 3
Parallel Form Reliability

	T-Value	Degrees of Freedom	2-Tail Probability
Prose prompts (between topics)	.10	10	.92
Graph prompts (between topics)	-.38	10	.71
Homelessness Topic (between prompt types)	.04	14	.97
Water Topic (between prompt types)	-.28	10	.79

Correlational Data

Correlational data from the reliability analysis was compared to determine whether students' relative standing on the test would be affected by the use of the two prompt types. A multiple correlations matrix was also run to determine the strength of the relationships between prompt types and subgroups of rhetorical scores (Table 4), and between the various evaluations of students' essays (Table 5). Correlations between prompt types and only rhetorical holistic scores are given in this report, since the sample size for the other three evaluations is too small to provide substantial statistical analyses. These data will be provided in future reports when data from an adequate sample size become available. Results relevant to the rhetorical holistic rating were prioritized in light of their direct application to the ELCE.

Effect on relative standing on test. Correlations were computed between the rhetorical scores for essays based on the prose prompts and the rhetorical scores for essays based on the graphic prompts. Correlations were then computed between rhetorical scores for essays within prompt types, essentially providing another measure of parallel form reliability. The extent to which correlations between scores across prompt types exceeded the correlations between scores within prompt type gives an indication of the strength of the influence of prompt type effects on students' relative standing on the test. The correlational data are shown in Table 2a above. The mean correlation for within prompt type correlations (.76) which is not included in the table was computed by taking the weighted average of transformed (z) scores and retransforming the average to an r statistic. The correlation between scores across prompt types (.76) and the correlation between scores within prompt types (.76) were essentiallyt identical. The later statistic serves as a statistic of parallel form reliability and effectively represents the highest possible correlation that could be expected for the relation between scores across prompt types. The fact that the correlations were nearly identical suggests that students' relative standing on the test was not significantly affected.

Prompt type x subgroup scores. Correlations between prompt types and scores assigned to various subgroups were analyzed to determine if there is a pattern of consistently similar scores associated with a particular prompt type for all students or for particular subgroups of students. The subgroups of interest are: 1) proficiency in writing English essays as defined by course level, 2) academic status, 3) field of study, 4) and native languages. Significant correlations between prompt type and scores would suggest that ratings are significantly affected by the particular prompt. If correlations across prompt types varied significantly, it could be argued that prompt type played a significant role in determining scores.

Correlational data for the prompt type x subgroup scores relationship is presented in Table 4 below, with data for proficiency levels (rows 1 - 8), academic status (rows 9 - 10), field of study (rows 11 - 12), and native languages (rows 13 - 17). Analysis reveals four statistically significant correlations, FW101-NS scores and FW111/112-1 scores with "water" prompts. Scores for Freshman Writing NSs' essays correlate with the "homelessness"-prose prompt at r (Pearson Correlation) = .59 ($p < .05$, $N = 17$) and with the "homelessness"-graphic prompt at $r = .65$ ($p < .05$, $N = 15$). However, the difference between the correlations is insignificant, suggesting that prompt type effects are insignificant. Similar implications can be drawn from the significant correlations between scores assigned to FW111/112-1 essays and the "water" prompts. These scores correlated with the "water"-prose prompt at $r = .69$ ($p < .05$, $N = 17$), with the "water"-graphic prompt at $r = .60$ ($p < .05$, $N = 22$). Again a comparison of these correlations yields an insignificant difference, suggesting insignificant prompt type effect. Although irrelevant to the issue of prompt type effect, the difference in FW111/112-1 scores across topics is interesting. The correlations of these scores to the "homelessness"-prose prompt at $r = .31$ ($p < .05$, $N = 16$) and

Table 4
Correlations between Scores and Subgroups

Subgroup	Prose Prompts		Graph Prompts	
	Water topic	Homelessness topic	Water topic	Homelessness topic
Proficiency levels[24]				
FW101-NS	-------	.59	-------	.65
FW101-NNS	-------	.24	-------	.32
FW111/112-1	.69	.31	.60	.20
FW111/112-2	-.20	-.36	-.11	-.25
ALI240-1	.26	-.05	.18	-.07
ALI240-2	-.38	-.29	-.38	-.36
ALI221	-.08	-.20	-.26	-.28
ALI220	-.34	-.27	-.36	-.32
Academic Status				
Undergraduate	.34	.31	.31	.42
Graduate	-.14	-.31	-.32	-.42
Field of Study				
Science / Engineering	.11	.02	-.07	-.14
Other	-.09	-.02	.07	.14
Native language[25]				
Chinese	-.04	-.23	.11	.23
Korean	.20	-.41	-.15	-.17
Japanese	.07	-.05	-.01	-.13
Indonesian	.16	-.04	.08	-.09
Other (Excluding English)	.02	.01	-.02	.01

to the "homelessness"-graphic prompt at $r = .20$ ($p < .05$, $N = 14$) contrast sharply with the correlations summarized above. Given the fact that no instructors used either of the topics represented in this study prior to collection of protocols, there is no apparent explanation for the discrepancy between correlations across topics.

[24] The proficiency level designations can be interpreted as follows:
FWP101-NS : Native English speaker Freshman Writing 101 students
FWP101-NNS : Non-native English speaker Freshman Writing 101students
FWP111/112-1 : Non-native English speaker Freshman Writing 111/112 and ALI 262 students in the upper 50% percentile as determined by this study only
FWP111/112-2 : Non-native English speaker Freshman Writing 111/112 and ALI 262 students in the lower 50% percentile as determined by this study only
ALI240-1 : Advanced ESL students in upper 50% percentile as determined by this study only
ALI240-2 : Advanced ESL students in lower 50% percentile as determined by this study only
ALI221 : Intermediate ESL science/technology graduate students
ALI220 : Intermediate ESL non-science/technology students

[25] Figures for native English speakers are equivalent to figures for FWP101-NS in Tables 4, 6a, 6b, 7a, and 7b. These figures are not reduplicated under the native language subgroup rubric. This accounts for the total sample size of 30 for the native language data subset in Tables 6a, 6b, 7a, and 7b.

Evaluation type correlations. An analysis of the correlations between the types of evaluations was undertaken to determine whether textual abstraction possibly played a significant role in determining raters' holitstic scores for rhetorical analysis. Pearson product-moment coefficients for pairs of evaluations are shown in Table 5 below. For each pair of evaluations, the Pearson *r* for scores associated with each prompt are given, then the mean correlations for scores associated with prompt types (i.e., prose or graph, without regard to topic), and then the mean correlation for scores for all essays.

The consistently high positive correlations between the two holistic ratings (rows 1 - 3 of Table 5) suggest that many of the same textual characteristics may be measured by both types of ratings. This would not be surprising according to Berthoff's (1986) view of textual abstraction as being inextricably bound to rhetoric. It is also possible that both ratings measure different textual characteristics which are equally salient in the essays. This possibility will be explored in a facet analysis yet to be completed. The use of an analytic scale would also possibly be useful in sorting out the various textual characteristics which influence an holistic rating but will require considerably more effort and is consequently beyond the immediate scope of this study. The primary source of these significantly high correlations is assumed to be raters' application of the scales. Discussions of the rhetorical ratings were recorded and will provide one source of research into the use of the rhetorical scale. Recording of the discussions of the application of the holistic scale of abstraction was not possible. However, as the other 382 essays are rated according to this scale, discussions among raters will be recorded for analysis. A comparitive study of the two sets of discussions may give insight into the apparent overlap between the two types of ratings.

The other correlational information is highly tentative at this point, given the small size of the subsample, and the non-normal distributions of the scores for the holistic evaluation of abstraction and, particularly, of the scores for the two MF / MD analyses. The generally significantly negative correlational statistics from this initial analysis, however, would suggest that the holistic ratings represent measurement of a different construct of abstraction than those of the two MF / MD analyses. This appears particularly true of the 'Informational-Involved' analysis which correlates significantly with the rhetorical ratings at $r = -.64$ ($\underline{p} < .05$, $\underline{N} = 30$) and with

Table 5

Correlations between Evaluations[26]

	Prose Prompts		Graph Prompts	
Evaluation	Water topic	Homelessness topic	Water topic	Homelessness topic
Rhetorical holistic with holistic evaluation of abstraction levels	.97	.63	.54	.85
Mean correlation between holistic evaluations for each prompt type	.89		.73	
Mean correlation between holistic evaluations for all essays	.83			
Rhetorical holistic with MF / MD evaluations	-.62 / -.47	-.02 / -.22	-.77 / -.18	-.34 / -.01
Mean correlation between rhetorical holistic and MF / MD evaluations for each prompt type	-.69 / -.36		-.60 / -.10	
Mean correlation between rhetorical holistic and MF / MD evaluations for all essays	-.64 / -.23			
Holistic evaluation of abstraction and MF / MD evaluations	-.77 / -.52	-.35 / .25	.03 / -.39	-.47 / .03
Mean correlation between evaluations of abstraction for each prompt type	-.60 / -.31		-.44 / -.36	
Mean correlation between evaluations of abstraction for all essays	-.53 / -.33			
MF / MD 'I-I' and 'A-N'[27] evaluations	.72	-.12	-.27	-.50
Mean correlation between MF / MD 'I-I' and 'A-N' evaluations for each prompt type	.38		-.39	
Mean correlation between MF / MD 'I-I' and 'A-N' evaluations for all essays	-.01			

[26] The first figure presented in rows 4-6 represents the correlation between holistic rhetorical scores and MF/MD 'Informational-Involved' scores; the second figure the correlation between holistic rhetorical scores and MF/MD 'Abstract-Nonabstract' scores. The first figure in rows 7-9 represents the correlation between holistic ratings of abstraction and MF/MD 'Informational-Involved' scores; the second figure the correlation between holistic ratings of abstraction and MF/MD 'Abstract-Nonabstract' scores.

[27] The 'Informational-Involved' evaluation is indicated as 'I-I'; the 'Abstract-Nonabstract' score as 'A-N'.

holistic ratings of abstraction at $r = -.53$ ($\underline{p} < .05$, $\underline{N} = 30$). In contrast, the two MF / MD textual analyses appear to be indepedent of each other, with a general correlation of $r = -.01$ ($\underline{p} < .05$, $\underline{N} = 30$). The analyses of a larger sample will provide more substantial results.

Performance Effects

Mean performance. In addition to analysis of relative ranking of students on the test, an analysis of effects on mean scores is also necessary due to the use of a criterion-referenced rating system. Mean scores of essays based on prose prompts and graphic prompts for subgroups are presented in Tables 6a and 6b. Table 6a presents data from the rhetorical holistic rating of all 412 essays. Data for the randomly chosen subsample of 30 essays from the holistic evaluation of abstraction levels is presented in Table 6b. A comparison of mean scores across prompt types for the rhetorical rating reveals that the differences in mean scores for essays are not significantly affected by prompt types. Not only are scores comparable across

Table 6a
Performance Data for Subgroups:
Holistic Rhetorical Rating for N = 412

	Prose	Prompt		Graph	Prompt	
	N	Mean	SD	N	Mean	SD
Proficiency levels						
FW101-NS	17	28.59	.93	15	29.07	.80
FW101-NNS	8	27.12	1.88	9	26.89	2.26
FW111/112-1	33	26.61	2.17	36	24.61	1.59
FW111/112-2	43	19.72	2.23	42	20.76	1.51
ALI240-1	42	22.40	1.46	46	22.33	1.50
ALI240-2	35	18.88	1.16	34	19.15	1.05
ALI221	12	19.17	2.33	15	19.00	2.45
ALI220	16	18.00	2.80	9	17.00	2.18
Academic status						
Undergraduate	134	22.48	4.18	132	23.05	3.53
Graduate	69	20.51	2.27	77	20.39	2.40
Field of study						
Science / Engineering	72	22.16	3.92	85	21.40	3.16
Other	131	21.76	3.78	124	22.42	3.56
Native language						
Chinese	92	21.07	2.94	105	21.66	2.76
Korean	31	18.90	3.20	30	20.87	3.16
Japanese	17	21.65	2.83	16	21.19	2.17
Indonesian	14	22.38	2.90	15	21.87	2.77
Other (Excluding English)	33	23.15	3.96	27	21.59	3.34

prompt types, but there is not consistent direction in the change across prompt types for all of the four major subgroups. A T-test analysis of scores indicates that, in fact, there is no statistically significant difference in scores across prompt type for any of the main subgroups (see Table 3 above). The differences observed between mean scores for the other three types of evaluations will be briefly discussed below.

Comparisons of mean scores for the holistic evaluation of abstraction (see Table 6b) across prompt types reveals significant differences, but these are very tentative results given the small sample size for each subgroup. The particularly notable difference in mean scores for the FW111/112-2 group is due to two very low scores. Again, the evaluation of a larger sample must be completed before any reliable conclusions can be drawn.

Table 6b
Performance Data for Subgroups:
Holistic Evaluation of Abstraction[28] for N = 30

	Prose Prompt			Graph Prompt		
	N	Mean	SD	N	Mean	SD
Proficiency levels						
FW101-NS	2	13.50	.50	2	21.00	50.00
FW101-NNS	2	21.50	12.50	2	22.50	6.75
FW111/112-1	4	16.75	51.58	4	14.75	12.92
FW111/112-2	4	5.75	3.58	4	12.25	26.25
ALI240-1	0	--------	--------	0	--------	--------
ALI240-2	2	10.50	12.50	0	--------	--------
ALI221	0	--------	--------	0	--------	--------
ALI220	1	4.00	0.00	3	6.00	4.67
Academic status						
Undergraduate	13	12.62	55.76	13	15.38	40.42
Graduate	2	10.50	12.50	2	6.50	12.50
Field of study						
Science / Engineering	5	13.40	32.06	3	16.67	41.33
Other	10	11.80	56.84	12	13.58	50.66
Native language						
Chinese	5	11.60	55.30	4	13.50	43.00
Korean	3	9.67	69.33	2	12.00	128.00
Japanese	1	4.00	0.00	2	9.00	32.00
Indonesian	1	6.00	0.00	2	14.50	12.50
Other (Excluding English)	3	20.33	.33	3	15.33	60.33

[28] The lack of observations for ALI 240-1, ALI 240-2 (graph prompt), and ALI 221 represented in rows 5-7 in Tables 5b, 6a and 6b is due to random sampling. No samples of these groups were obtained in the 30 essays that were randomly chosen from the original 412 essays. To provide for adequate statistical analyses, a larger, more representative sample will be included as work on the study continues.

Likewise, there are significant differences between mean scores from the MF / MD analyses, but these must be considered very tentative. Mean MF / MD scores and standard deviations for principle subgroups are summarized in Tables 7a and 7b below.

Table 7a
Data for Subgroups:
MF / MD Analysis along Dimension 1
'Informational vs. Involved Production' for N = 30

	Prose	Prompt		Graph	Prompt	
	N	Mean	SD	N	Mean	SD
Proficiency levels						
FW101-NS	2	41.65	254.93	2	18.46	62.50
FW101-NNS	2	22.12	45.12	2	33.22	264.50
FW111/112-1	4	23.52	41.49	4	29.18	18.91
FW111/112-2	4	48.60	362.29	4	43.38	508.75
ALI240-1	0	-------	-------	0	-------	-------
ALI240-2	2	30.86	508.75	0	-------	-------
ALI221	0	-------	-------	0	-------	-------
ALI220	1	50.19	0.00	3	36.75	54.98
Academic status						
Undergraduate	13	36.91	17.18	13	33.20	15.40
Graduate	2	30.86	7.62	2	36.08	10.36
Field of study						
Science / Engineering	5	39.97	19.02	3	28.73	6.73
Other	10	32.81	14.54	12	34.80	15.92
Native language						
Chinese	5	33.54	16.90	4	36.61	23.41
Korean	3	46.92	59.11	2	34.25	18.32
Japanese	1	50.19	0.00	2	56.18	36.20
Indonesian	1	23.26	0.00	2	30.98	8.33
Other (Excluding English)	3	20.90	13.58	3	25.88	21.16

Table 7b
Data for Subgroups:
MF / MD Analysis along Dimension 5
'Abstract vs. Non-Abstract Information' for N = 30

	Prose	Prompt		Graph	Prompt	
	N	Mean	SD	N	Mean	SD
Proficiency levels						
FW101-NS	2	1.95	4.15	2	8.17	4.15
FW101-NNS	2	3.89	3.92	2	1.66	10.90
FW111/112-1	4	3.76	13.38	4	2.09	17.80
FW111/112-2	4	4.74	8.33	4	1.98	5.58
ALI240-1	0	--------	--------	0	--------	--------
ALI240-2	2	4.91	6.99	0	--------	--------
ALI221	0	--------	--------	0	--------	--------
ALI220	1	3.10	0.00	3	8.99	60.85
Academic status						
Undergraduate	13	3.76	2.32	13	2.91	3.61
Graduate	2	4.92	2.62	2	12.56	6.72
Field of study						
Science / Engineering	5	6.08	1.10	3	4.15	4.27
Other	10	2.80	2.04	12	4.12	5.22
Native language						
Chinese	5	4.15	4.96	4	2.15	8.46
Korean	3	5.77	1.45	2	8.66	17.29
Japanese	1	3.10	0.00	2	1.58	.57
Indonesian	1	.47	0.00	2	2.90	5.86
Other (Excluding English)	3	4.38	7.93	3	3.90	7.76

Principle statistical analyses

Three omnibus statistical procedures were completed to determine the existence of a significant prompt effect as a determining factor of scores assigned to students' essays. First, in order to identify variables which played a significant role in determining the variance in scores, the SAS procedure, GLM, with a stepwise regression was run. Because of the unbalanced research design reflected in Table 1, the GLM was necessary to obtain an analysis of variance, since it compensates for unequal representation of variables through analysis of Sums of Squares (SS) unlike the ANOVA. The GLM provides statistical evidence useful in determining the adequacy of all chosen variables to account for variance in the dependent variable. The stepwise regression was used to identify the particular independent variables significant in accounting for variance in students' scores.

Results obtained from the stepwise regression were useful in the discriminant analyses which were used determine whether scores could be distinguished according to assigned prompt type. The variables identified in the regression as being most effective in determining overall variance in scores were used to determine if particular subgroups of students were more influenced by prompt type assignment than were others.

Stepwise discriminant analysis combines stepwise regression analysis and discriminant analysis procedures to identify independent variables most significant in determining group membership as a function of the amount of variance accounted for by each independent variable. This final omnibus procedure identified sources of variance associated specifically with prompt type effects, as opposed to overall variance as in simple stepwise regression analysis.

Model adequacy analysis. In addition to a regression analysis, the GLM procedure provides an indication of the adequacy of the choice of independent variables (referred to as the model) in accounting for variance. This information is highlighted here due to the fact that the adequacy of the model lays a foundation for the reliability of subsequent interpretations of statistical results. Ideally, a researcher would hope that the model would account for 100% of the variance in the dependent variable. This would indicate that the factors which interact to determine variance in the dependent variable have all been identified. One could assume under these circumstances that no other factors were significant in determining the behavior of the dependent variable. However, this ideal situation rarely occurs in research involving human characteristics as variables although researchers attempt to account for as much variance i" the data as possible.

The GLM procedure was run to determine the success of the model to account for the variation in students' rhetorical scores. The regression analysis, with rhetorical scores as the dependent variable and 40 independent variables[29], proved effective with an R-squared of 80% and was statistically significant ($\underline{F} = 38.16$, $\underline{df} = 39$, $\underline{p} < .0001$). The choice of independent variables, then, accounted for 80% of the variance in scores. This provides a fairly solid basis for the reliability of interpretations of further statistical results.

[29] The 40 independent variables used in the GLM are: 1) prompt type, 2) topic, 3) academic status, 4) field of study, 5) 14 variables defined by native language, 6) 8 variables defined by proficiency level, 7) 4 variables defined by number of years of ESL/EFL instruction, 8) 4 variables defined by number of years of English essay writing experience, 9) fluency in writing native language, 10) 4 variables defined by age, and 11) sex.

Stepwise regression analysis. Stepwise regressions identify the independent variables which are significant in accounting for the variance in the independent variable. Students' rhetorical scores were again used as the dependent variable with the same 40 independent variables used in the GLM. The regression analysis for this study was run under the GLM. The significant results for the regression are summarized in Table 8 below. Academic status (i.e., undergraduate versus graduate), two native languages, and five of the proficiency level variables appear to be significant in determining the variations in students' rhetorical scores. Notably, prompt type did not prove to be a statistically significant variable in predicting variation in students' scores. Prompt type, in fact, appears insignificant ($F = .06$, $p < .80$), with a Type III SS of 0.18. Topic also appears relatively insignificant ($F = 0.99$, $p < .32$), with a Type III SS of 2.84.

Three of the student characteristics mentioned in the introduction as possible promoters of prompt type effect, thus, appear significant in determining the overall variance in students' rhetorical scores. It was suggested that academic status and proficiency level might distinguish students who successfully address graphic prompts from those who do not. It was also hypothesized that various language groups may also be more successful than others. Because variables defined by languages were inconsistent in significantly accounting for variance in scores (i.e., only 2 out of 14 are significant)[30] , and because they proved insignificant as factors in determining group membership according to prompt type in the discriminant analyses, these variables will not be further discussed.

Because proficiency levels were fairly consistent in accounting for variance, these variables and other variables, including academic status, were pursued in the discriminant analyses summarized below (Table 9).

[30] Two other variables associated with native languages approached significance in the stepwise regression. Both Chinese ($F = 2.97$, $p < .08$) with a Type III SS of 8.58, and Thai ($F = 3.08$, $p < .08$) with a Type III SS of 8.87 neared significance.

Table 8
Stepwise Regression Results

Independent variable	Type III SS	F-value	Probability > F
Academic status	36.40	12.63	.0004
Korean	18.35	6.37	.01
English	35.38	12.27	.0005
FW101-NNS	554.24	192.26	.0001
FW111/112-1	726.17	251.91	.0001
FW111/112-2	129.04	44.76	.0001
ALI240-1	278.96	96.77	.0001
ALI240-2	23.13	8.02	.005

The GLM procedure was also run on data supplied by additional analyses of the subsample of 3C essays with scores of holistic evaluations of abstraction, 'Informational-Involved' scores, and "Abstract-Nonabstract' scores as dependent variables. While these statistical analyses are very tentative in light of the small sample size, it should be noted that prompt type did not appear significant in determining overall variance in scores representing holistic ratings of abstraction or in either of the two sets of quantitative scores.[31]

Discriminant analyses. A number of discriminant analyses were performed in order to determine which subgroups of students might be most affected by prompt type effects. In the discriminant analysis, the independent variable was group membership defined by prompt type; the same 40 independent variables were entered as in the GLM analyses. The first goal in performing the discriminant analyses was to determine if the scores, or subsets of scores, could be distinguished according to assigned prompt type. The second goal in running the discriminant analyses was to identify the combination of variables that will make the best discrimination.

A summary of the results is given in Table 9, with indications of the subgroup of data used in each analysis, total sample size for the subgroup and the type of scores, and the percentage of scores correctly classified. The first row in the summary represents the discriminant analysis of rhetorical scores for all 412 essays. Rows 2 - 8 provide data for various subgroups and types of scores. Because the first goal in running each of the analyses was not met (i.e, scores were not

[31] Type III SS was not calculated due to the small N size. An analysis of Type II SS, however, indicates that prompt type accounts for less than 1% of the observed variance in each of these sets of scores.

discriminated according to prompt type assignment), statistics for individual variables are not summarized.

The rhetorical scores of native English speakers were most accurately distinguished according to prompt type (with 64% accuracy). This suggests that prompt type effects are more predominant in native speakers' writing. This may be due partially to native speakers not being distracted as much by concerns about language. They may be able to devote more attention to the use of graphics in compensation for less attention to concerns about language. There appears to be a slight effect from prompt type assignment on science / engineering majors also (classified with 58% accuracy) as hypothesized. The figure for undergraduate students (58% accuracy) is skewed by the inclusion of native speakers' scores. Prompt type effects appear to be more influential when essays are rated using the holistic evaluation of abstraction (classified with 63% accuracy). This particular observation must be tentative due to the small sample size and questions about the application of the scale outlined above. All of the observations above are very tentative due to the low rate of classification accuracy. In light of 80% accuracy being the threshold for significance in discriminant analyses, analyses summarized can, at best, be considered as approaching significance.

Table 9
Discriminant Analyses Results

Subgroup	# / Type of Cases	% of Cases Correctly Classified
All cases	412 Rhetorical Scores	51
Undergraduates	267 Rhetorical Scores	58
Graduates	145 Rhetorical Scores	52
NSs	32 Rhetorical Scores	64
Science Majors	158 Rhetorical Scores	58
Subsample	30 Holistic Scores of Abstraction	63
Subsample	30 'Informational-Involved' Scores	52
Subsample	30 'Abstract-Nonabstract' Scores	48

Stepwise discriminant analysis. The dependent variable in the stepwise discriminant analysis was group membership defined by prompt type; the same 40 independent variables were used in this analysis as in all other omnibus anslyses. The goal in running the stepwise discriminant analysis was to determine which, if any, variables were significant in discriminating scores as a function of the amount of variance in scores for which they accounted. Instead of identifying groups of variables which discriminate as in a simple discriminant analysis, a stepwise discriminant analysis checks the amount of variance associated with each variable as a factor of its power to discriminate. Consonant with earlier findings, none of the 40 variables was statistically significant in discriminating scores according to prompt type assignment.

Questionnaire and Interview Data

Responses from both questionnaires are currently being tabulated; transcription of all interviews must also be completed. The administration of questionnaires and interviews was intended to provide data relevant to students' reactions to the use graphics on a test of writing abilities. Additionly, interpretations of the specific prompts, including graphics, was sought. These types of information may be helpful not only in defining sources of variance in scores, but serve a useful purpose in their own right, providing information about the "face validity" of including graphics on a standardized composition test from students' viewpoint. Chi-square analyses will be used to compare results from questionnaires for statistical significance in distinguishing prompt type assignment. Interviews will be analyzed qualitatively for additional insight in prompt type effects. Both sets of data will be helpful in completing an analysis of test method and expected response facets according to Bachman's framework (Bachman 1990, Bachman and Palmer, forthcoming).

Preliminary analyses of the interview data are inconclusive. Interviewees generally indicate a great deal of familiarity with the "homelessness" topic while the "water" topic appears to be less familiar. This may account for some of the variation in responses across topics and is consistent with regression results which suggest that topic effects are greater than prompt type effects although still not statistically significant. All science / engineering students indicated that they were very familiar with graphics and that they felt comfortable with the use of graphics on essay prompts. In contrast, some humanities students had difficulty interpreting the graphics and expressed discomfort with the incorporation of graphics in essay prompts. Finally, while students assigned one of the prose prompts often complained about the lack of information provided in the prompt, students assigned graphic prompts rarely expressed the same concern.

Discussion

All statistical analyses summarized in this report suggest that the use of both prose and graphic essay prompts does not significantly reduce the reliable measurement of students' writing abilities as defined in specifications for the writing section of the ELSE. Mean scores across prompt types did not significantly differ for any proficiency level or any other subgroups of students defined by academic status, field of study, and native language. Parallel form reliability and rater correlations indicate no significant inconsistency in ratings due to prompt type assignment. Furthermore, tentative analyses of the various evaluations of essays suggest that there are no significant differences in textual complexity due to prompt type assignment, accounting for at least one textual characteristic which could affect raters' decisions. However, students' responses to questionnaires and interview questions indicate some concern about the face validity of including graphics in prompts destined for testing purposes. Analyses of textual abstraction and of students' responses to questionnaires and interview questions are tentative at this time, awaiting further progress of the research study. Analyses of textual abstraction will become more substantial with the evaluation of a larger sample of essays. Students' responses will be analyzed for statistical significance and will serve as a useful source of information in an analysis of test method characteristics.

In many respects, then, the character of the test appears to be unchanged with the addition of the graphic prompt type. Nevertheless, both statistical evidence and student reactions to test methods should be considered in addressing the question regarding the use of both prompt types for an essay test such as the ELCE.

There are a number of apparent plausible explanations for the results which suggest that prompt type assignment has an insignificant influence on students' performance on composition tests. One possibility is that the composition of the FWP and ALI international student populations masked the differences which might become apparent with a population more representative of the foreign student population in the United States. It may be that the heavy representation of ethnic Chinese (48% of the sample) or the heavy representation of Asians (at least 78% of the sample) in general was a significant factor in determining the results of the study. As Carlson *et al*. (1985) suggest, the significance of prompt type effects may vary from one student population to another.

It would be useful, then, to replicate this study with a comparable student population with a different ethnic composition.

A more likely explanation is the nature of the rhetorical ratings which may have minimized the significance of an essay's "content" and the discourse features associated with incorporating data from graphics or other outside sources of information into one's own discourse effectively. Discussions among the first set of raters, who completed the rhetorical ratings, about the relative importance of "content," "organization" and "grammar" seemed to end with the conclusion that sentence level grammar should be prioritized. If it is true that sentence level grammar was the primary focus of the raters, it would not be surprising that the differences that graphics might promote in the "content" and "organization" (i.e., the two aspects most likely to be affected) of students' essays would not be significant in determining scores. Transcriptions of raters' discussions may be useful in defining the relative significance of the various textual characteristics assigned by raters in scoring essays.

Another possible factor which may interact with the two factors identified above, but which may likely be the most significant one, is the nature of students' construal of the task presented by the prompts. During the piloting of the materials and the completion of MF / MD analyses, the researcher noted that students' uses of the information incorporated in the graphics varied greatly. Some students, for example, all but ignored the graphics, pulling only linguistically coded information from the rubrics of a limited number of graphics. In contrast, some students made extensive use of both linguistically and numerically coded information which they culled from all three graphics. The vagueness of the instructions regarding the use of graphics in the graphic prompts allowed for this discrepency in task construal. The instructions are vague in the sense that students were not told to use a specific type or amount of information from the graphics. Some essays based on the graphic prompts, in fact, can not be identified according prompt type assignment without reference to the coded information indicating the assignment. Moreover, since the incorporation of specific data was not a scoring criterion, and since raters' main focus may have been on sentence level grammar, prompt type assignment may have had no influence on scoring. This has an obvious confounding effect on the ability to distinguish essays according to prompt type assignment, especially if there is a significant number of such essays. Facet analyses, as outlined below, will be useful in teasing out this information. Also, it may be that once evaluations of abstraction in a larger sample are completed, prompt type assignment will appear significant. Data from the prompt evaluation forms may also prove useful in this regard.

Suggestions for Further Research

A number of lines of related research have been identified in this report as useful to the study of prompt type effects. The most immediate need is for facet analyses of test method and expected responses represented in the prompts. Descriptions of the characteristics of the prompts and of expected reponses should be compared to determine whether the prompts provided adequate context for the responses expected by the researcher. Similarly, a comparative study of the expected response facet and students' actual responses should be completed with the goal of determining to what extent students' writing was constrained by the writing prompts. Data from recorded interviews and from the two questionnaires completed by students may be helpful in this regard.

Although the model appeared efficient in accounting for most of the variance in the rhetorical scores, model "fit" will be further analyzed through the use of the SAS (1989) procedure, structural equation modeling. This procedure will provide a means for testing various models' effectiveness in identifying sources of variance in scores.

Another line of research underlies the construction of the holistic scale of abstraction and the comparison of the types of evaluations. Although rhetorical scales often include the descriptor, "content," as one criterion against which essays are scored, it is not clear what is meant by this descriptor. The term is often used in opposition to other descriptors, "organization," "development," "grammar," and "mechanics" (Bridgeman and Carlson 1983, Freedman 1977, 1979, Horowitz 1986). It is variably used to refer to knowledge schemata (e.g., the "originality" of ideas, the cogency of arguments, or the general use of "supporting evidence"), to affective schemata (e.g., the apparent stance of the writer with regard to an issue), and to language schemtata (e.g., the use of discourse markers). The construction of a scale to measure the abstraction of argumenation is an attempt not only to determine the influence of textual abstraction on raters' decisions, but also to tease out the vague notion of "content" as a textual characteristic.

It is important to point out that results from this study and Carlson *et al* .'s (1985) study can not be used to suggest that Bridgeman and Carlson's (1983) descriptions of the two prompt types are invalid. The two prompt types may, in fact, engender different tasks, requiring the employment of varying sets of cognitive abilities. These differences may be masked, however, by a number of factors, including the compensating effects of holistic rating. This issue is beyond the immediate scope of the present study but represents a line of research critical to the analysis of prompt type effects and test method effects in general.

Concluding Remarks

All statistical analyses summarized in this report suggest that prompt type effect is insignificant as a predictor of student performance on tests of writing ability. In spite of preliminary data from student interviews which indicate some concern about the validity of using graphics on tests of composition, there is no statistical evidence that indicates that prompt type significantly affects students' performance. We may conclude, then, that the two prompt types described in this report may be used on the ELCE with assurance that doing so will not significantly affect the reliability of raters' scores. This should be particularly true should the test taker population be composed largely of science / engineering students. The validity of these conclusions will, however, remain tentative pending completion of the study.

References

Adler, R. R. 1971. An investigation of the factors which affect the quality of essays by advanced placement students. Urbana, IL: University of Illinois. Ph.D. Dissertation.

Bachman, L. F. 1990. *Fundamental considerations in language testing* . Oxford: Oxford University Press.

Bachman, L. F. and A. S. Palmer. forthcoming. *Language testing in practice*.. Oxford: Oxford University Press.

Baker, E. L. 1982. The specification of writing tasks. In A. Purves and S. Takala (eds.) *An International perspective on the evaluation of written composition* . Oxford: Pergamon. 291-297.

Bereiter, C. and M. Scardamalia. 1987. *The psychology of written composition*.. Hillside, NJ: Lawrence Erlbaum Associates.

Bereiter, C. and M. Scardamalia. 1983. Levels of inquiry in writing research. In P. Mosehthal, L. Tamor, and S. Walmsley (eds.) *Research on writing: Principles and methods* . New York: Longman. 3 - 25.

Berthoff, A. E. 1986. Abstraction as a speculative instrument. In D. A. McQuade (ed.) *The territory of language: Linguistics, stylistics, and the teaching of composition.*. Carbondale, IL: Southern Illinois University Press. 227-237.

Biber, D. 1988. *Variation across speech and writing* . Cambridge: Cambridge University Press.

Bridgeman, B. and S. B. Carlson. 1983. *Survey of academic writing tasks required of graduate and undergraduate foreign students.* (TOEFL Research Report No. 15.) Princeton, NJ: Educational Testing Service.

Brossell, G. 1983. Rhetorical specification in essay examination topics. *College English.*. 45. 165-173.

Brossell, G. and B. Ash. 1984. An experiment with the wording of essay topics. *College composition and communication.*. 35. 423-425.

Carlson, S. B., B. Bridgeman, R. Camp, and J. Waanders. 1985. *Relationship of admission test scores to writing performance of native and nonnative speakers of English.*. (TOEFL Research Report No. 19.) Princeton, NJ: Educational Testing Service.

Carroll, J. B. 1961. Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English proficiency of foreign students* . Washington, D. C.: Center for Applied Linguistics. 30-40.

Celce-Murcia, M. 1991. Discourse analysis and grammar instruction. In W. Grabe *et al* . (eds.) *Annual review of applied linguistics, 11*. New York: Cambridge University Press. 135-151.

Connor, U. 1987. Argumentative patterns in student essays: Cross-cultural differences. In U. Connor and R. B. Kaplan (eds.) *Writing across languages: Analysis of L2 text.*. Reading, MA: Addison-Wesley. 57-71.

Delia, J., S. Kline and B. Burlson. 1979. The development of persuasive communication strategies in kindergarteners through twelfth-graders. *Communication monographs* . 46. 241-256.

Duranti, A. 1985. Sociocultural dimension of discourse. In T. A. Van Dijk (ed.) *Handbook of discourse analysis: Volume 1 Disciplines of discourse*. New York: Academic Press. 193-230.

Educational Testing Service. 1986. *Test of Written English (TWE) scoring guidelines* . Princeton, NJ: Educational Testing Service.

Educational Testing Service. 1989. *Test of Written English (TWE) guide* . Princeton, NJ: Educational Testing Service.

Ferris, D. 1990. Linguistic and rhetorical characteristics of student argumentative writing by native and non-native speakers of English. Los Angeles: University of Southern California. Qualifying paper.

Flower, L. and J. Hayes. 1977. Problem-solving strategies and the writing process. *College English.*. 39. 449-461.

Flower, L. and J. Hayes. 1980. The cognition of discovery: Defining a rhetorical problem. *College composition and communication.*. 31. 21-32.

Freedman, A. and I. Pringle. 1980. Writing in the college years: Some indices of growth. *College composition and communication..* 31. 311-324.

Freedman, S. W. 1977. Influences on the evaluation of student writing. Los Angeles, CA: University of California at Los Angeles. Ph.D. Dissertation.

Freedman, S. W. 1979. How characteristics of student essays influence teachers' evaluations. *Journal of educational psychology* . 71. 3. 328-338.

Goodwin, C. and A. Duranti. forthcoming. Rethinking context: An introduction. In C. Goodwin and A. Duranti (eds.) *Rethinking context* . Ms.

Greenberg, K. L. 1982. Some relationships between writing assignments and students' writing performance. *The writing instructor* . 2. 7-14.

Greenberg, K. L. 1986. The development and validation of the TOEFL writing test: A discussion of TOEFL Research Reports 15 and 19. *TESOL Quarterly* . 20. 3. 531-544.

Hake, R. and J. M. Williams. 1981. Style and its consequences: Do as I do, not as I say. *College English..* 43. 5. 433-451.

Hale, G. A. 1991. Effects of amount of time allowed on the TOEFL Test of Written English. Draft of final report submitted to TOEFL Research Committee.

Hambleton, R. K., H. Swaminathan, and H. J. Rogers. 1991. *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications.

Hayward, M. 1989. Choosing an essay test question: It's more than what you know. *Teaching English in the two-year college* . 16. 174-178.

Hayward, M. 1990. Evaluations of essay prompts by nonnative speakers of English. *TESOL Quarterly* . 24. 4. 753-757.

Hoetker, J. 1982. Essay examination topics and students' writing. *College composition and communication..* 33. 377-392.

Horowitz, D. 1986. What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly* . 20. 3. 445-462.

Kaplan, R. B. 1966. Cultural thought patterns in intercultural education. *Language learning* . 16. (Reprinted in K. Croft (ed.) 1972. *Readings on English as a second language for teachers and teacher trainees* . Cambridge, MA: Winthrop. 2nd ed. 1980.)

Keech, C. L. 1984. Apparent regression in student writing performance as a function of unrecognized changes in task complexity. Berkley, CA: University of California at Berkley. Ph.D. Dissertation.

Lado, R. 1961. *Language testing* . New York: McGraw-Hill.

Moustafa, M. 1987. The effects of word processing on composing in a university ESL composition class. Los Angeles: University of Southern California. MS.

Nelson, J. 1990. *This was an easy assignment: Examining how students interpret academic writing tasks* . (Technical Report No. 43.) Berkley, CA: University of California at Berkley, Center for the Study of Writing.

Odell, L. 1981. Defining and assessing competence in writing. In C. R. Cooper (ed.) *The nature and measurement of competency in English..* Urbana, IL: National Council of Teachers of English. 95-138.

Oller, J. W. Jr. 1979. *Language tests at school: A pragmatic approach* . London: Longman.

Park, Y. and A. Purves. None. The influence of the task upon writing performance in English as a second language. MS.

Peck, W. C. 1989. *The effects of prompts upon revision: A glimpse of the gap between planning and performance* . (Reading-to-write Report No. 7.) Berkley, CA: University of California at Berkley, Center for the Study of Writing.

Perkins, K. 1983. One the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly* . 17. 4. 651-671.

Poole, D. 1990. Contextualizing IRE in an eighth grade quiz review. *Linguistics and Education.* 2. 185-211.

Quellmalz, E., F. Capell, and C. Chou. 1982. Effects of discourse and response mode on the measurement of writing competence. *Journal of educational measurement.* 19. 241-258.

Purcell, E. T. and R. W. Suter. 1980. Predictors of pronunciation accuracy: A reexamination. *Language learning* . 30. 2. 271-287.

Raimes, A. 1990. The TOEFL Test of Written English: Causes for concern. *TESOL Quarterly* . 24. 3. 427-442.

Ruth, L. 1982. Sources of knowledge for designing writing test prompts. In J. R. Gray and L. Ruth (eds.) *Properties of writing tasks: A study of alternative procedures for holistic writing assessment* . Berkley, CA: University of California at Berkley, Graduate School of Education, Bay Area Writing Project. 31-131. [ERIC No. ED 230 576.].

Ruth, L. and S. Murphy. 1988. *Designing writing tasks for the assessment of writing* . Norwood, NJ: Ablex.

Skehan, P. 1990. Progress in language testing: The 1990s. In C. Alderson and B. North (eds.) *Language testing in the 1990s* . London: MacMillan. 3-21.

Statistical Analysis System Institute. 1989. *Statistical Analysis System, version 6.0* . Cary, NC: SAS.

Statistical Package for the Social Sciences. 1990. *Statistical Package for the Social Sciences, version 4.1* . Chicago: SPSS.

Suter, R. W. 1976. Predictors of pronunciation accuracy in second language learning. *Language learning* . 26. 233-253.

Williams, J. M. 1979. Defining complexity. *College English* . 40. 6. 595-609.

Winters, L. 1979. The effects of differing response criteria on the assessment of writing competence. Los Angeles, CA: University of California at Los Angeles. Ph. D. Dissertation.

APPENDIX 1

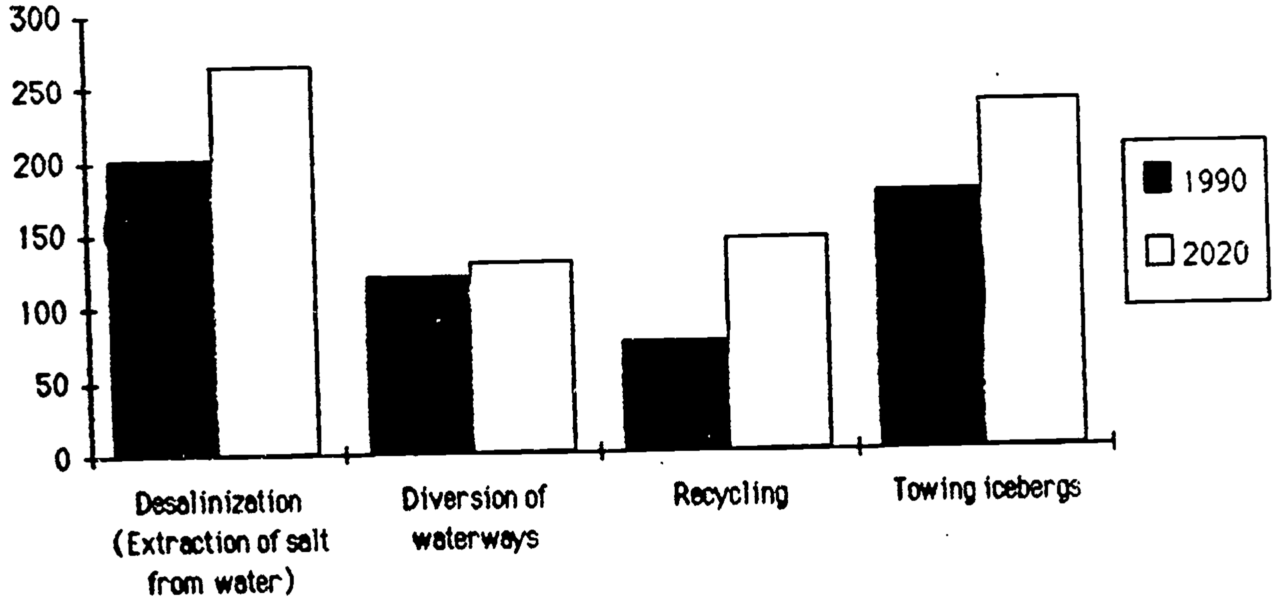
Essay Prompts

Prompt 1A   Student # ____________

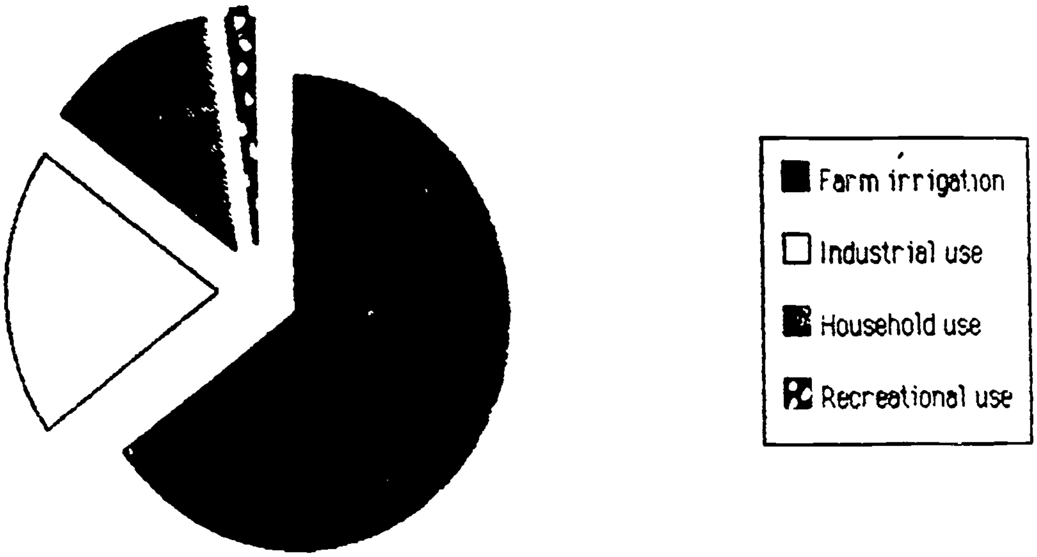
____

____

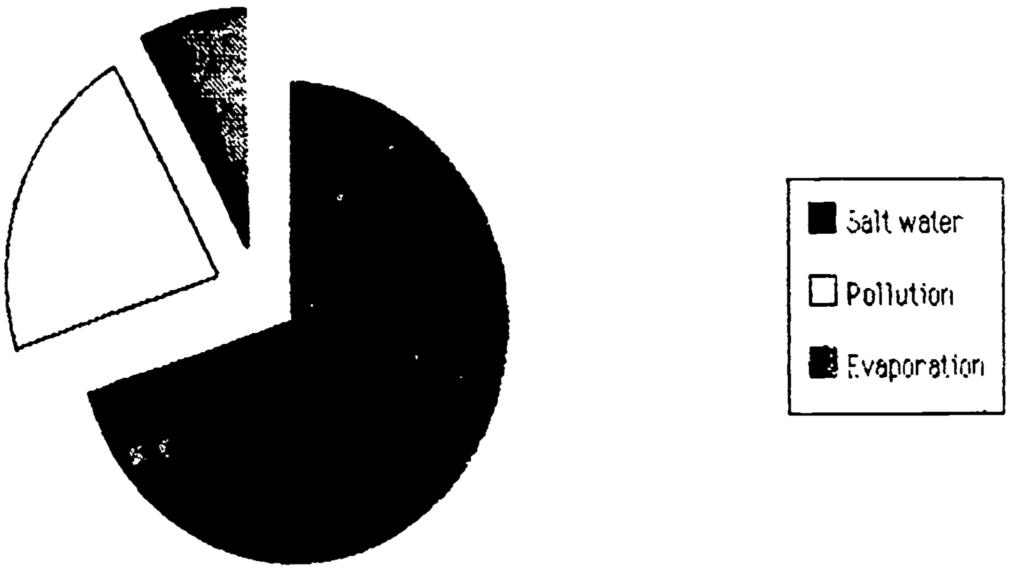
The global supply of fresh water is quickly decreasing due to population growth, increased pollution and overuse of the supply. Propose solutions to solve the problem, using information from the graph and charts to support your argument.



Number of U.S. Dollars to produce 1 liter of fresh water



Percentages of all water used, by use, 1990 figures



Percentage of all unused water, by cause, 1990 figures

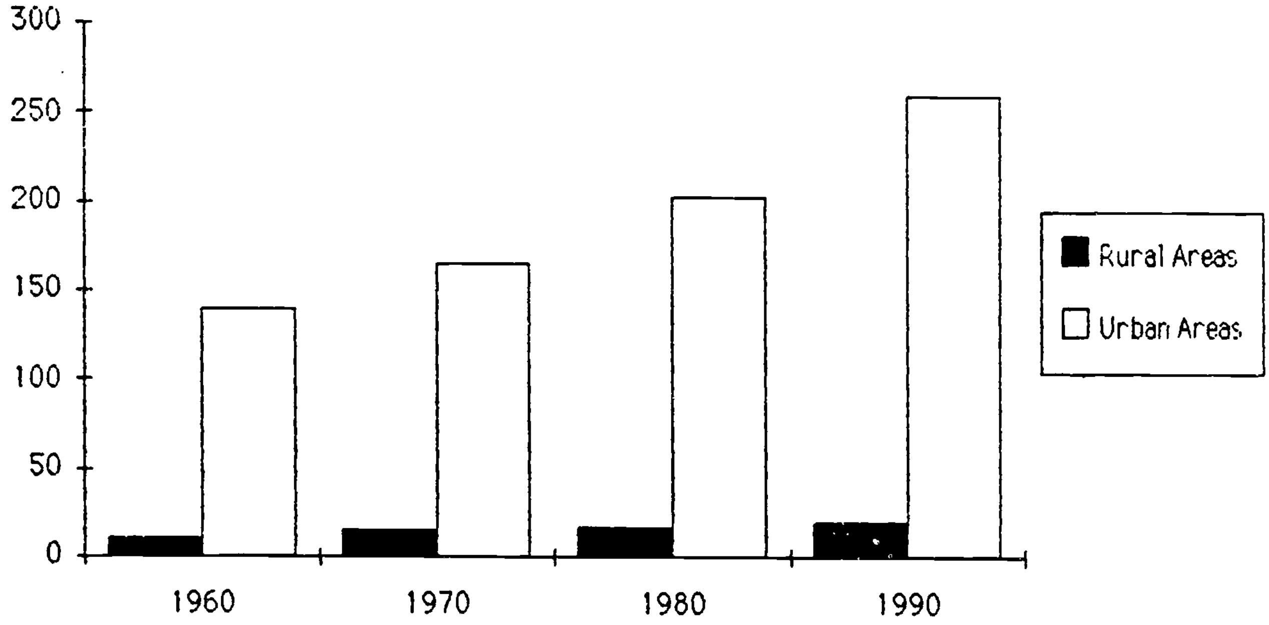
Prompt 2A

Student # ______________

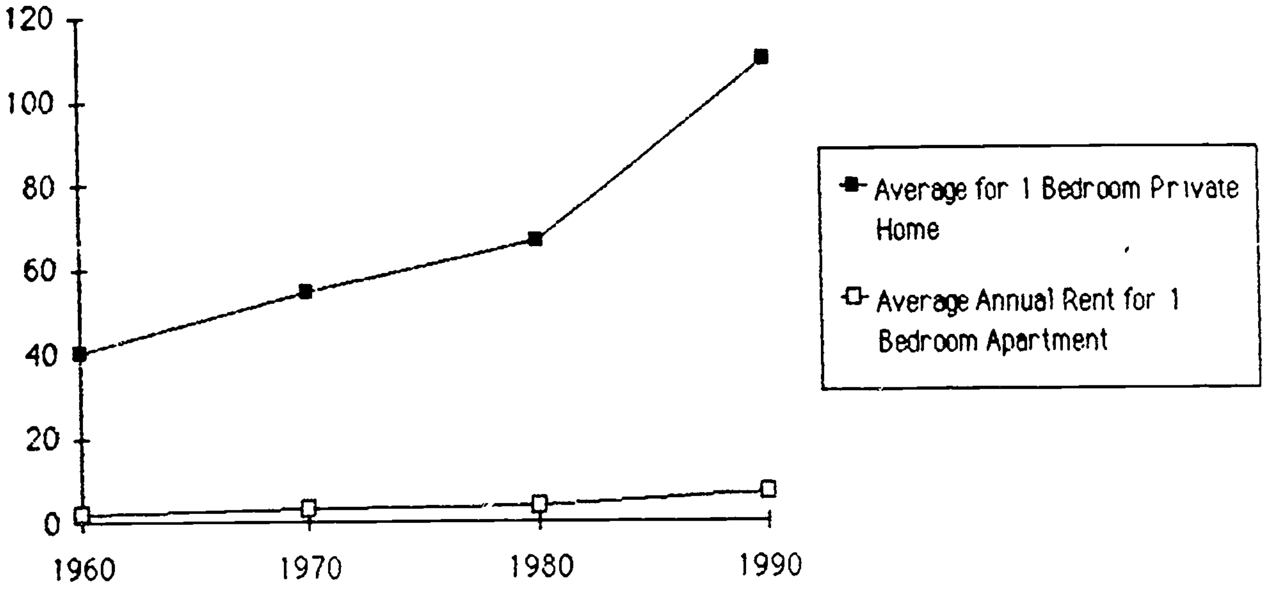
____

____

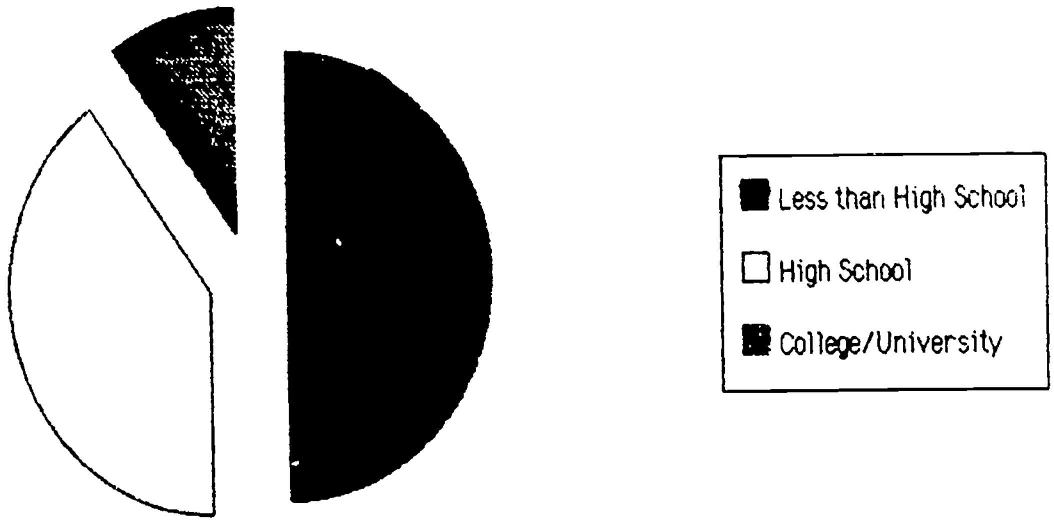
Homelessness is becoming a major problem around the world due to slowing economies, rising costs of living and changes in the types of skills demanded of employees. Propose solutions to solve the problem, using information from the graphs and chart to support your argument.



World's homeless population, by area (in millions)



Average costs for home and annual apartment rent, by year (in thousands of U.S Dollars), 1990 U.S. figures



Highest education level completed by homeless, by percentage of total homeless population, 1990 U.S. figures

Prompt 1B   Student # ____________

____

____

The global supply of fresh water is quickly decreasing due to population growth, increased pollution and overuse of the supply. Propose solutions to solve the problem and support your argument.

Prompt 2B Student # ______________

____

____

Homelessness is becoming a major problem around the world due to slowing economies, rising costs of living and changes in the types of skills demanded of employees. Propose solutions to solve the problem and support your argument.

APPENDIX 2

Questionnaires

Questionnaire 1

Student # ________
Prompt # ________

Native language: ____________________ Number of years of ESL/EFL study: _____ years

Are you a fluent writer in your native language? Yes No (Circle one) Age: ______ years

Number of years of experience writing essays in English: _____ years Sex: M F (Circle one)

Status: Undergrad Grad (Circle one) Major/Department: ______________________

Nationality: __________________

1. I read the prompt _______ times before I felt that I fully understood it.

2. Circle the words or phrases of the prompt which you consider the most important (that clarify the writing task for you).

3. Compared to other topics I have written about, this one is ... (Choose one in each line.)

Easy	1	2	3	4	5	Difficult
Boring	1	2	3	4	5	Interesting
Familiar	1	2	3	4	5	Unfamiliar
Unclear	1	2	3	4	5	Clear
Challenging	1	2	3	4	5	Simple
Abstract	1	2	3	4	5	Meaningful

4. The prompt ... (Check all that apply.)

___ is too long / ___ is too short ___ is interesting / ___ is boring

___ uses simple vocabulary / ___ uses difficult vocabulary

___ clearly explains the topic / ___ does NOT clearly explain the topic

___ presents a topic I know A LOT about / ___ presents a topic I know LITTLE about

___ will help me develop ideas for writing / ___ will NOT help me develop ideas for writing

___ will give me information to support my arguments / ___ will NOT give me information to support my arguments

OVER

5. **If your prompt includes graphics**, please answer the following question. If your prompt does NOT include graphics, please do NOT answer this question.

The graphics ... (Check all that apply.)

___ are simple / ___ are complex

___ are easy to understand / ___ are difficult to understand

___ are relevant to the topic / ___ are NOT relevant to the topic

___ are related to each other / ___ are NOT related to each other

___ will help me develop ideas for writing / ___ will NOT help me develop ideas for writing

___ will give me information to support my argument / ___ will NOT give me information to support my argument

___ There are TOO MANY graphics. / ___ There are TOO FEW graphics.

7. **If your prompt included graphics**, please complete this question. If your prompt did NOT include graphics, please go to the next question.

The graphics ... (Check all that apply.)

___ were easy to understand / ___ were difficult to understand

___ were relevant to the topic / ___ were NOT relevant to the topic

___ were related to each other / ___ were NOT related to each other

___ helped me develop ideas for writing / ___ did NOT help me develop ideas for writing

___ gave me information to support my arguments / ___ did NOT give me information to support my arguments

8. Rank order the concerns that you had while writing. (1=most concerned about this, 2=next most concerned about this, etc.). Rank only those concerns that actually occurred to you; omit those items that do not reflect your concerns.

___ ideas

___ spelling/punctuation

___ purpose for writing

___ clear statement of point of view

___ use of formal / informal English

___ coherence (The essay forms a meaningful whole text.)

___ cohesion (The relations between words and thoughts are clearly expressed.)

___ organization

___ audience

___ grammar

___ revisions

9. Compared to other essays you have written, this one is ...

Poorly written 1 2 3 4 5 Very well written

Questionnaire 2

Student # __________
Prompt # __________

1. Reconstruct the prompt below as accurately as possible.

____

____

____

____

2. Did you read the entire prompt before writing? ___ Yes ___ No

Yes, I read it ... ___ once ___ three times
(Check only one.)
___ twice ___ more than three times

3. Did you reread the prompt while writing? ___ Yes ___ No

Yes, I reread it ... ___ once ___ three times
(Check only one.)
___ twice ___ more than three times

4. How many minutes did you spend deciding what to write before actually writing your essay?

_________ minutes

5. Did you have difficulty deciding what to write about? ___ Yes ___ No

6. The prompt ... (Check all that apply.)

___ was easy to understand / ___ was difficult to understand

___ presented a topic I know A LOT about / ___ presented a topic I know LITTLE about

___ helped me develop ideas for writing / ___ did NOT help me develop ideas for writing

___ gave me information to support my arguments / ___ did NOT give me information to support my arguments

OVER

APPENDIX 3

Interview Questions

Interview 1 questions Student #______

---

1. Do you feel that you fully understand the prompt?

2. What is the prompt asking? Give a verbal response to the prompt.

3. Is any part of the prompt unclear or difficult to read? If so, which? Can you explain why?

4. What part of the prompt makes it easy/difficult to understand?

5. Do you think that this prompt is asking for writing that is similar to other↲
←writing you've done? How so/Why not?

6. Who is going to be the audience? What role will you as writer take?
←What is the purpose for writing? What mode of development will you
←use?

7. What parts (words, clauses, sentences, etc.) of the prompt do you think are↲
←the most important? Why?

8. Will this be a difficult writing assignment? Why/Why not?

9. Please complete the following sentences.

This prompt would have been clearer if ....

←This prompt would have been easier to read if ....

Interview 2 questions

Student # ____

1. Do you feel that you fully understood the prompt?

2. What did the prompt ask? Give a verbal response to the prompt.

3. Was any part of the prompt unclear or diffucult to read? If so, which?
Can you explain why?

4. What part of the prompt made it easy/difficult to understand?

5. Did this prompt ask for writing similar to writing you have done in the past? How so?/Why not?

6. Who did you assume to be your audience? What was your role as writer?
What was your purpose in writing? What mode of development did you use?

7. What parts (words, clauses, sentences, etc.) of the prompt did you think were the most important? Why?

8. Was this a difficult writing assignment? Why/Why not?

9. Please complete the following sentences.

The prompt would have been clearer if ...

The prompt would have been easier to read if ...

APPENDIX 4

Prompt Evaluation Form

Prompt Analysis Project
Initial Prompt Evaluation Form

Name ______________________
Don Weasenforth

Please evaluate the attached prompt by circling the number in each line below which best represents your opinion. Your comments after each section are encouraged. The last page of the questionnaire is provided for the continuation of your comments. This prompt will be given to students in ALI 220, 221, 240 and 262 as well as to non-native English speaking students in Freshman Writing.

I. Topic

Appropriate student sample	1	2	3	4	**Inappropriate** for for student sample
Inaccessible to proficiency range	1	2	3	4	**Accessible** to proficiency range
Inaccessible to various disciplines	1	2	3	4	**Accessible** to various disciplines
Relevant to students' concerns	1	2	3	4	**Irrelevant** to students' concerns
Too **broad**	1	2	3	4	Too **constraining**
Boring	1	2	3	4	Interesting
Prone to **cliches** or stereotypical thinking	1	2	3	4	Will lead to original thinking
Threatening to personal values	1	2	3	4	**Nonthreatening** to personal values
Nonthreatening to religious beliefs	1	2	3	4	**Threatening** to religious beliefs
Nonthreatening to self esteem	1	2	3	4	**Threatening** to self esteem
Socially **biased**	1	2	3	4	Socially **neutral**

(continued on back)

I. Topic (continued)

Culturally **neutral**	1	2	3	4	Culturally **biased**
Clearly expressed	1	2	3	4	**Unclear**
Challenging	1	2	3	4	Simple
Abstract	1	2	3	4	Meaningful

Comments:____

____

____

__________________________________(continued on last sheet)

II. Statement of topic

Too **elaborate**	1	2	3	4	Too **concise**
Topic is **clearly** defined	1	2	3	4	Topic is **vaguely** defined
Single topic assigned	1	2	3	4	**Multiple** (implicit) topics possible
Scope of argument **beyond** students' competence	1	2	3	4	Scope of argument **within** students' competence
Restricts writer to **one** viewpoint	1	2	3	4	Allows for viable **choice** of opinion
Concepts do **not** require specialized knowledge	1	2	3	4	Concepts require specialized knowledge

II. Statement of topic (continued)

Vocabulary **is** accessible to range of proficiencies	1	2	3	4	Vocabulary is **not** accessible to range of proficiencies
Syntax used is **not** accessible to range of proficiencies	1	2	3	4	Syntax used **is** accessible to range of proficiencies

Comments: ______________________________

____

____

______________________________(continued on last sheet)

III. Instructions (Description of task)

Too **concise**	1	2	3	4	Too **elaborate**
Too **directive**	1	2	3	4	Too **vague**
Clearly states expected use of graphic(s)	1	2	3	4	Does **not** clearly state expected use of graphic(s)
Expected mode of discourse **is** clearly stated	1	2	3	4	Expected mode of discourse **not** clearly stated
Concepts require specialized knowledge	1	2	3	4	Concepts do **not** require specialized knowledge
Expected level of specificity is clearly stated	1	2	3	4	Expected level of specificity is **not** clearly stated

(continued on back)

III. Instructions (continued)

Vocabulary is **not** accessible to range of proficiencies	1	2	3	4	Vocabulary **is** accessible to range of proficiencies
Syntax used **is** accessible to range of proficiencies	1	2	3	4	Syntax used in **not** accessible to range of proficiencies

Comments:

____

____

____

______________________________(continued on last sheet)

IV. Graphic(s)

Visually **intimidating**	1	2	3	4	Visually **agreeable**
Print is **legible**	1	2	3	4	Print is **illegible**
Difficult to interpret	1	2	3	4	**Easy** to interpret
All data **is** relevant to topic	1	2	3	4	Data is **not** relevant to topic
Contain(s) culturally **biased** data	1	2	3	4	Contain(s) culturally **neutral** data
Contain(s) socially **neutral** data	1	2	3	4	Contain(s) socially **biased** data
Is/Are in culturally **biased** format	1	2	3	4	Is/Are in culturally **neutral** format
Data **will** bias writer's viewpoint	1	2	3	4	Data will **not** bias writer's viewpoint

IV. Graphic(s) (continued)

Usefulness of data to task is **obvious**	1	2	3	4	Usefulness of data to task is **obscure**
Relationship between graphics is **obscure**	1	2	3	4	Relationship between graphics is **obvious**

Comments: ____

____

____

__________________________________(continued on last sheet)

APPENDIX 5

Traditional Rhetorical Holistic Rating Scale

ESL CHALLENGE TEST
ESSAY EVALUATION SCALE

1-2 Incompetence	3-4 Minimal Competence
*Indistinguishable introduction, body, conclusion *No apparent thesis	*Apparent introduction, body, conclusion *Inexplicit thesis
*Off-topic, incoherent response *Irrelevant, unconnected details within paragraph(s); no focus *No logical development; paragraphs not related to one another; don't develop a thesis *No or inappropriately used cohesive devices	*On topic; may not address all elements of the question *Insufficient, sometimes irrelevant, detail within paragraph(s); no focus *Inadequate development; paragraphs usually related, but do not adequately or appropriately develop a thesis *Few cohesive devices, sometimes inappropriately used
*Severe problems with word forms *Vocabulary range extremely limited *Frequent errors in word choice, often affecting intelligibility *Frequent grammatical errors, affecting intelligibility *Range of syntactic structures extremely limited; short simple sentences	*Many problems with word forms *Very limited vocabulary range *Frequent errors in word choice, occasionally affecting intelligibility *Frequent grammatical errors, occasionally affecting intelligibility *Range of syntactic structures very limited; some compound but no complex sentences
*Paragraphing conventions often violated *Punctuation inappropriate, inconsistent *Capitalization inappropriate, inconsistent *Frequent spelling errors, often affecting intelligibility	*Paragraphing conventions sometimes violated *Punctuation sometimes inappropriate, inconsistent *Capitalization sometimes inappropriate, inconsistent *Frequent spelling errors, occasionally affecting intelligibility

5-6 Developing Competence	7-8 Proficiency w/ Some Errors
*Explicit introduction, body, conclusion; may lack unity	*Clear introduction, body, conclusion; may not be perfectly unified
*Explicit but poorly formulated thesis	*Clear but weak thesis
*Addresses all elements of question, but not adequately	*Addresses all elements of question; some parts may be slighted
*Insufficient paragraph development; inconsistent use of topice sentences; inadequate detail; paragraphs sometimes not focused	*Clear paragraph development; topic sentences and supporting detail; clear focus
*Paragraphs related, but may not adequately or appropriately develop a thesis	*Paragraphs related, but may not adequately or appropriately develop a thesis
*Rudimentary cohesive devices	*Some variety in cohesive devices
*Occasional problems with word forms	*Very few word form problems
*Limited vocabulary	*Fairly broad sub-technical vocabulary
*Some minor word choice errors; some register problems	*Very few word choice errors; register appropriate
*Occasional grammatical errors, rarely affecting intelligibility	*Very few grammatical errors
*Syntactic structures somewhat varied; some complex sentences	*Variety of syntactic structures; some complex sentences
*Paragraphing conventions observed consistently	*Paragraphing conventions observed consistently
*Punctuation sometimes inappropriate or inconsistent	*Punctuation appropriate consistent
*Capitalization appropriate and consistent	*Capitalization appropriate and consistent
*Occasional spelling errors	*Infrequent spelling errors

9-10 Near Native Proficiency

*Clear, unified introduction, body, conclusion
*Clear, effective thesis

*Comprehensive response to the prompt
*Good paragraph development; topic sentences and ample detail; clear focus
*Paragraphs related; adequately and appropriately develop the thesis
*Variety of cohesive devices

*Correct word forms
*Broad, sophisticated vocabulary
*Hardly any word choice errors; register appropriate
*Almost no grammatical errors
*Syntactic structures varied; many complex sentences

*Paragraphing conventions used consistently
*Punctuation appropriate
*Capitalization appropriate
*Almost no spelling errors

APPENDIX 6

Holistic Rating Scale for Levels of Abstraction

Levels of Abstraction/Generalization

Criteria:
1. Level of abstraction of information from graphics
2. Level of generalization evident in incorporation of information from graphics in text
3. Text awareness
4. Distance from topic
5. Function of text (expected to be held constant)
6. Linguistic features
7. Level of moral development

Level — Description/Example

I

1. No apparent use of information from graphics or minimal superficial information used in reference to graphics;no use numerical values from graphs;focus on similar features (e.g., on dates to exclusion of other features) (e.g., Hydro energy is used less./ Oil and gas is used most./ Nuclear energy will increase from 1975 to 2020.)
2. No information from graphics incorporated in text or minimal information repeated which may not coherently support argument; if data used, mostly in description of graphics
3. Text is wholly or greatly comprised of writer's personal experience which may not address topic; text may be mostly narrative and in present tense with heavy reliance on organization by coordination

7. Appeals to established authority, fixed concrete rules and the maintenance of social order for its own sake; correct behavior is the accomplishment of one's duty

II

1. Abstraction of information from graphics based on comparison of information directly encoded by graphics (i.e., comparisons of information within one graphic and especially among different graphics; focus still on similarities (e.g., Solar, coal and nuclear sources will be equally used in 2000.); abstraction at lower level may also be used
2. Relatively little information from graphics used to support arguments; text may give impression of forced manipulation to accommodate information from graphics; reliance on description of data instead of use to support arguments; parts of argumentation (proposition, arguments,

evidence) apparent, but do not form complete whole

3. Writer recounts personal experience or experience which seems to have direct influence on writer which may not directly address topic; text may contain some narrative; text may be mostly recounted in past tense with some reliance on organization by coordination; irrational use of categorical statement (using universal quantifiers, such as all, never, only) without specific support; argument by analogy dominant; strict tripartite organizational structure

6. Failed attempts to use transition phrases to mark logical development

III

1. Abstraction of information beyond that explicitly encoded in graphics or statement of topic (e.g. extrapolation of past

and/or

future trends, or interpolation of intermediary points of graphics); abstraction at lower levels may also be used

2. Writer shows good understanding of graphics; writer shows ability to correlate related information within and among graphics; writer shows ability to infer information from graphics, to speculate on past and future dates and intermediary values of the graphics; abstracted information coherently supports arguments most of time; text is mostly driven by arguments, not by need to incorporate data from graphics although effective use of data is apparent

3. Writer may rely on comparison and/or contrast; mixed reliance on coordination and subordination for organization; laborious overstatement of slight ideas; cumbersome iteration or repetition of points; many transition words/phrases used but some used inappropriately; some summary used after representational writing showing attendance to audience needs; argument discourse format apparent, but some parts more developed than others (e.g., clearly stated proposition, but few arguments or little evidence or some arguments well supported while others not); linking of ideas from disparate levels of abstractness

7. Orientation toward social contract; right behavior is determined by general individual rights and standards agreed upon by society; emphasis on procedural rules for consensus and on legal point of view and on possibility of changing rules

IV

1. Information abstracted by mathematical processing (e.g., ratios, percentages) of explicit data is used (e.g., Use of hydro energy sources in 2020 will be 1/3 of use of nuclear sources in the same year.)
2. Arguments are supported by information representing a range of abstraction, including mathematical manipulation of data explicitly stated in graphics; text is driven by purpose to convince or persuade, not need to incorporate data although effective use of data is evident
3. Organization marked by subordination; view of issue incorporates society/humanity; argument schema apparent and parts are all adequately developed, including counter-argument; solutions offered in place of summary conclusion

V

1. Coordination of data from all graphs at various levels of abstraction into coherent whole theoretical view of issue/ abstraction, synthesis and generalization of data in coherent theory
2. Use of data and interpolations and/or extrapolations coherently support arguments
3. Adaptation of formal argument schema; hierarchical organization of propositions; proficiency evident although formal conventions of argumentation not necessarily adhered to
6. Evidence of metadiscourse strategies (e.g., effective summary and elaboration)

7. Appeal to universal ethical principles; rights decided by conscience in accord with own logical, ethical principles, not concrete moral rules; emphasis on principles of justice, equality of human rights and respect for dignity of human beings as individual persons